Estrategia computacional para detección y caracterización de bloques microsíntenicos relacionados a regiones genómicas asociadas a domesticación en frijol Lima

Leidy Tatiana García Navarrete

Universidad Nacional de Colombia
Facultad de Ingeniería, Departamento de Ingeniería de sistemas e industrial
Bogotá, Colombia
2018

Estrategia computacional para detección y caracterización de bloques microsíntenicos relacionados a regiones genómicas asociadas a domesticación en frijol Lima

Leidy Tatiana García Navarrete

Tesis presentada como requisito parcial para optar al título de:
Magister en Bioinformática

Directora:
(Ph.D.) María Isabel Chacón Sánchez

Co-Directora:
(Ph.D.) Clara Isabel Bermúdez Santana

Grupos de Investigación:
RNómica teórica y computacional
Horticultura

Universidad Nacional de Colombia
Facultad de Ingeniería, Departamento de Ingeniería de Sistemas e Industrial
Bogotá, Colombia
2018

Dedicatoria

A mis padres, Luz Marina y Emigdio, por su amor y paciencia durantes estos años. A mis hermanos, Jimy y Oscar, por su apoyo incondicional, y a mis sobrinos, por su ternura que le regalan a mi vida.

"Estoy convencido de que la mitad de lo que separa a los emprendedores exitosos de los que no triunfan, es la perseverancia."

Steve Jobs

Agradecimientos

Agradezco a mi directora, la profesora María Isabel Chacón por sus múltiples enseñanzas sobre la genómica del frijol Lima, por sus consejos y apoyo incondicional que permitieron la culminación de este proyecto.

Al profesor Jorge Duitama, por su colaboración e invaluables aportes en la presente investigación.

A la profesora Clara Bermúdez por sus orientaciones y dedicación al desarrollo del presente trabajo.

A Paola Hurtado y Laura Torres, por su apoyo constante en las largas jornadas de trabajo de campo y sus enseñanzas en el arte del laboratorio de biología molecular.

A mis compañeros del grupo de RNómica teórica y computacional, en especial a Valentina, Oscar, Aimer, Gabriel, Jeny y Astrid.

A mis amigos: Sandra Sáyer, Daniel Tello, Juanita Gil, Jair Alarcón, Stephany Rueda y Ángela Caicedo.

A la Facultad de Ciencias Agrarias y la Vicerrectoría de Investigación de la Universidad Nacional de Colombia, y a Colciencias por los fondos para la realización de este trabajo, bajo el marco del proyecto: Estructura genética y adaptación climática del frijol Lima *Phaseolus lunatus* L. y su domesticación: un nuevo enfoque mediante huellas genómicas obtenidas por secuenciación, número de contrato FP44842-009-2015 y código de proyecto 1101-658-42502.

Al servicio de intercambio Alemán-DAAD y a la Facultad de Ciencias de la Universidad Nacional de Colombia por el programa de subvensión y equipamiento del Laboratorio de Biologia Computacional en el cual se realizaron parte de los análsis de esta tésis.

A la Dirección de Servicios de Información y Tecnología (DSIT) de la Universidad de los Andes, por el soporte técnico e infraestructura de cómputo de alto rendimiento, y en especial al doctor Juan Pablo Mallarino.

Al Centro Internacional de Agricultura Tropical (CIAT), en especial al doctor Daniel De-

bouck y al doctor Peter Wenzl, jefe de la unidad de recursos genéticos. A las investigadoras Adriana Bohórquez, Eliana Macea y María Eugenia Recío, por su asesoría en protocolos de extracción de mRNA.

Resumen

El frijol Lima (*Phaseolus lunatus* L.) es la segunda especie domesticada más importante del género *Phaseolus* y es de interés desde el punto de vista agronómico y ecológico por el amplio rango de adaptaciones agro-ecológicas que presenta. A pesar de esto, la carencia de recursos genómicos en frijol Lima ha sido uno de los mayores obstáculos para maximizar su potencial como cultivo o como fuente de rasgos de interés agronómico. En las leguminosas cultivadas, la dehiscencia de la vaina previo a la cosecha es indeseable puesto que causa grandes pérdidas en el rendimiento, por lo tanto, el conocimiento de los genes que controlan éste y otros rasgos importantes puede favorecer futuros programas de mejoramiento en esta especie. Para dar respuesta a esta necesidad, en el presente estudio se secuenció el genoma de una variedad domesticada de frijol Lima de Colombia con una combinación de plataformas de secuenciación (Illumina, PacBio y 10X-Genomics). El secuenciamiento produjo un total de 97.6 Gb de datos crudos, que después del filtrado y control de calidad generaron 61.9 Gb (103x de profundidad) que se usaron para el ensamblaje. Se obtuvo un ensamblaje genómico con una longitud de 541 Mpb, contenidos en 496 contigs, con un N50 de 5.5 Mpb. La longitud del ensamblaje representó el 90 % del tamaño estimado del genoma. Para la anotación funcional de este ensamblaje, se secuenciaron y ensamblaron los transcriptomas de tres tejidos vegetales (flor, hoja y vaina) que permitieron la identificación de 48.127 genes. Finalmente, el ensamblaje genómico se usó en un enfoque de genómica comparativa que detectó regiones microsinténicas entre frijol Lima, frijol común y frijol mungo. Esta estrategia permitió la identificación en frijol Lima de genes candidatos asociados a la dehiscencia de la vaina. Los resultados de la presente investigación son un avance significativo en la generación de recursos genómicos en frijol Lima que permitirán a la comunidad científica avanzar no solo en el estudio y mejoramiento genético de esta especie, sino también en un mayor entendimiento de la evolución de los cultivos de leguminosas en general.

Abstract

Lima bean (*Phaseolus lunatus* L.) is the second most important domesticated species of the genus *Phaseolus* and its wide range of agro-ecological adaptations makes it a species of scientific interest for agronomic and ecological research. In spite of this, the lack of genomic resources in Lima bean has been one of the major hurdles to maximize its potential as a crop or as a source of traits of agronomic interest. In crop legumes, pod dehiscence prior to harvest is an undesirable trait associated to yield loss, therefore knowledge of genes that control this and other important traits may favor future breeding programs. To address this urgent need, in this study the genome of a Colombian domesticated variety of Lima bean was sequenced using a combination of platforms (Illumina, PacBio and 10X-Genomics). We generated about 97.6 Gb of raw sequencing data that after cleaning and filtering yielded a total of 61.9 Gb (103x depth) that were used for assembly. A genomic assembly was ob-

tained with a total length of 541 Mbp, contained in 496 contigs, and an N50 of 5.5 Mbp. This genomic assembly was around 90 % of estimated genome size. To annotate this assembly, transcriptome data were obtained from three tissues (flower, leaf and pod) and used to identify 48.127 genes. Finally, the genome assembly was used in a comparative genomics framework to detect microsynthenic regions between Lima bean, common bean and mungo bean that contained candidate genes associated to pod dehiscence. Present results are a significant progress in the development of genomic resources in Lima bean, which will benefit not only the scientific community interested in the genetic improvement of this species but also researchers interested in understanding legume evolution in general.

Keywords: Ensamblaje genómico *de novo*, tecnologías de secuenciación de segunda y tercera generación, transcriptómica, RNA-seq, anotación funcional, genómica comparativa, leguminosa, dehiscencia de la vaina, *Phaseolus*.

Contenido

Lista de Figuras

Lista de Tablas

1. Introducción

Uno de los campos científicos más fascinantes y fundamentales es la biología evolutiva. Las especies evolucionan a través de la acumulación de cambios genéticos que ocurren a lo largo de su historia, cambios que se originan principalmente por procesos de mutación génica, que alteran los genomas de las especies a una pequeña escala, y mutaciones cromosómicas, que alteran tanto el número de cromosomas como la morfología de los mismos. La historia evolutiva de las especies se puede por lo tanto leer en sus genomas. En este sentido, descifrar y comparar el contenido genómico de las especies provee una herramienta muy poderosa para entender los principios y mecanismos que gobiernan los procesos biológicos.

La genómica comparativa, es decir, la comparación de genomas completos entre taxa y también de secuencias al interior de un genoma, nos permite entender los eventos que han direccionado la evolución de los genomas y sus genes, y por ende de las especies. Las estructuras genómicas que comparten dos o más especies y que descienden de un ancestro común se definen como homólogas. Los cambios genómicos pueden ser en sitios puntuales en la secuencia de ADN, secuencias de genes completos o bloques cromosómicos. Al ser adquiridas desde un ancestro común, las estructuras homólogas serán muy similares en secuencia y posiblemente en función.

Los genes que comparten dos especies y que presentan similitudes en secuencia se definen como genes homólogos y aquí se distinguen dos clases: los genes ortólogos son genes derivados de un evento de especiación y los genes parálogos son genes que se originaron en un genoma por eventos de duplicación [64]. Esta clasificación de ortología y paralogía se puede extender así mismo a nivel genómico, por lo tanto elementos genómicos que estaban presentes en el último ancestro común y que comparten las especies derivadas se definen también como ortólogos [64]. De las comparaciones a nivel genómico se ha observado que no sólo los genes pueden ser homólogos entre las especies sino también aquellos bloques de genes que retienen el mismo orden del ancestro común. A estos bloques se les conoce como bloques sinténicos.

En genomas recientemente secuenciados, la ortología de genes se infiere inicialmente por similaridad de secuencias y raramente se toma en cuenta la posición de los genes en el genoma. Los genes ortólogos pueden también ser definidos con base en la sintenia, es decir, si dos copias de genes presentes en dos especies relacionadas están acompañados por genes sinténicos adicionales, este patrón de microsintenia provee fuerte evidencia de las relaciones

de ortología entre los genes evaluados.

La evidencia experimental ha mostrado que los genes ortólogos tienden a retener funciones equivalentes entre las especies, aunque con excepciones [64], por lo que es necesario validar experimentalmente su función. En estudios comparativos la identificación de genes ortólogos permite realizar de manera eficiente la anotación funcional de un nuevo genoma. La anotación funcional a su vez facilita la predicción de genes que pueden estar involucrados en el control genético de un rasgo fenotípico de interés. Aunque la predicción de función por ortología no es completamente confiable, realizar esta predicción ha contribuido enormemente a seleccionar genes para experimentos de validación funcional en muchas especies. Los genes seleccionados para validación experimental se conocen como genes candidatos y la aproximación de detección de gen candidato ha sido muy importante para el estudio del control genético de muchos rasgos de interés. Por ejemplo, algunos rasgos que despiertan mayor interés en el estudio de las plantas cultivadas son aquellos relacionados con la domesticación debido a que muchos de ellos, como la dehiscencia de la vaina en las leguminosas, están directamente relacionados con el rendimiento en estas especies. Por lo anterior, es de interés conocer si estos rasgos se encuentran en regiones genómicas conservadas entre especies relacionadas.

Los estudios a nivel microsinténico o a escala de genoma completo (macrosintenia) [113], dependen grandemente de la disponibilidad de ensamblajes de genomas de alta calidad, con su respectiva anotación funcional y estructural, que permita la comparación entre especies. De acuerdo a lo anterior, en la era de la genómica el punto de partida para un estudio comparativo debe ser la generación de un ensamblaje genómico de la especie de interés (si éste no está disponible), seguido de su anotación estructural y funcional.

En la presente investigación, la especie de interés es la especie leguminosa conocida como frijol Lima (*Phaseolus lunatus* L.), una de las cinco especies domesticadas del género *Phaseolus* [31]. Esta especie es de interés desde el punto de vista agronómico debido a que a lo largo de su rango de distribución, que va desde el norte de México hasta el norte de Argentina, presenta un amplio rango de adaptaciones agro-ecológicas que hacen de ésta una especie modelo para el estudio de las adaptaciones biológicas [31]. El rasgo de interés en el presente estudio y para el cual se plantea la identificación de genes ortólogos a través de la genómica comparativa es la dehiscencia de la vaina. En las leguminosas, la dehiscencia del fruto es un rasgo indeseable debido a que su presencia puede causar grandes pérdidas en rendimiento del cultivo, por lo tanto el conocimiento de su control genético puede favorecer futuros programas de mejoramiento en esta especie.

Es por esto que el objetivo de la presente investigación fue generar una estrategia computacional que permitiera detectar en frijol Lima regiones microsinténicas asociadas a genes

de la domesticación, específicamente a genes relacionados con la dehiscencia de la vaina. En el presente estudio, primero se generó un ensamblaje *de novo* del genoma de un individuo de frijol Lima cultivado, posteriormente se ensambló el transcriptoma proveniente de tres tejidos diferentes como herramienta de apoyo para la anotación funcional del genoma, y finalmente se detectaron bloques microsinténicos entre frijol Lima y otras especies leguminosas asociados a genes de la domesticación, en específico al rasgo dehiscencia de la vaina. Los datos genómicos fueron obtenidos de una combinación de plataformas de secuenciación (Illumina, PacBio y 10X-Genomics) y los datos transcriptómicos (ARN mensajero) de tres tejidos diferentes (flor, hoja y vaina) fueron obtenidos de la plataforma Illumina.

La estrategia computacional inició con la evaluación de múltiples métodos de ensamblaje *de novo* del genoma para determinar cuáles producían los mejores ensamblajes. Las estrategias de ensamblaje tuvieron en cuenta la estructura de los datos (longitud de las lecturas, tasa de error, profundidad y cobertura) y las ventajas y limitaciones propias de las diferentes plataformas de secuenciación. Posteriormente, el ensamblaje más robusto fue anotado funcionalmente. Para este fin, se ensamblaron los transcriptomas provenientes de tres tejidos vegetales. El ensamblaje de los transcriptomas consideró tanto la estrategia *de novo* como el guiado por genoma de referencia. Los ensamblajes de transcriptoma fueron evaluados y empleados en la anotación del genoma de frijol Lima que consideró tanto la evidencia experimental como predicciones *ab initio*. Finalmente, se detectaron regiones microsinténicas asociadas al rasgo dehiscencia de la vaina entre frijol Lima y dos especies relacionadas de leguminosas (frijol común y frijol mungo). Se establecieron relaciones de microsintenia en regiones sub-genómicas que contenían cada uno de cinco genes que controlan la dehiscencia de la vaina (según lo reportado para la especie vegetal modelo *Arabidopsis thaliana*), y posteriormente los bloques microsinténicos detectados fueron caracterizados con relación al contenido génico y al grado de ordenamiento y orientación de los genes allí contenidos.

Debido a que la identificación de regiones microsinténicas involucra un proceso continuo desde el ensamblaje *de novo* del genoma, su anotación y posterior comparación de los ensamblajes obtenidos con secuencias de genomas de otras especies relacionadas, entonces el presente documento se ha dividido en cinco capítulos para una mejor comprensión. En el primer capítulo (el actual), se provee al lector con una introducción general sobre la naturaleza y alcance de la presente investigación. El segundo capítulo provee el marco conceptual y estado del arte de la presente investigación. El tercer capítulo reporta los resultados del ensamblaje *de novo* de alta calidad de un individuo de frijol Lima. El cuarto capítulo presenta los resultados del ensamblaje del transcriptoma de tres tejidos diferentes y la anotación estructural y funcional del genoma. El quinto capítulo presenta los resultados de la detección de bloques microsinténicos asociados al rasgo dehiscencia de la vaina. El documento finaliza con unas conclusiones y recomendaciones generales.

Los resultados de la presente investigación son un avance significativo en la generación de recursos genómicos en una especie vegetal de importancia económica, los cuales permitirán a la comunidad científica avanzar no solo en el estudio y mejoramiento del frijol Lima, sino también en un mayor entendimiento de la evolución de los cultivos de leguminosas en general.

2. Marco conceptual

2.1. ¿Qué es un Genoma?

El genoma se asume como "el grupo completo de secuencias del material genético de un organismo. Incluye la secuencia de cada cromosoma, más cualquier ADN presente en organelos como las mitocondrias y en las plantas los cloroplastos" [15], su expresión requiere de la acción coordinada de enzimas y otras proteínas que dan como producto inicial el transcriptoma y como segundo producto el proteoma. La obtención de la secuencia de bases de nucleótidos completa para todos los cromosomas de una especie, ha sido posible a través de los avances en las tecnologías de secuenciación, junto con el desarrollo de software bioinformático.

En el caso del genoma vegetal, éste se caracteriza por tener gran tamaño y poseer regiones de alta repetición, y en algunas especies por contener múltiples copias de cromosomas enteros (poliploidía) [26]. El genoma nuclear de una planta se encuentra dividido en cromosomas, con una longitud característica para cada tipo de especie, al igual que la composición de nucleótidos de las regiones repetitivas y de las regiones de copia única. En cuanto a los genes de las plantas, éstos tienen generalmente mayor contenido de guanina y citosina en los exones y contenidos más bajos en intrones [55]. Es importante resaltar que el conocimiento del tamaño y características de un genoma es un factor determinante para la selección de la estrategia de secuenciamiento y ensamblaje [52].

2.2. Fundamentos de las tecnologías de secuenciamiento de ADN

Determinar el orden de los nucleótidos en las moléculas de ADN es el primer paso para la caracterización genómica de una especie; para lograr ésto diversas tecnologías de secuenciamiento y de biología molecular han sido desarrolladas en los últimos cincuenta años. De acuerdo con las características de las plataformas de secuenciamiento se distinguen tres generaciones. La tecnología de primera generación se fundamentan en el método Sanger, donde originalmente se realizaban cuatro reacciones diferentes de síntesis de ADN, empleando un didesoxinucleótido (ddNTP), marcado con 32P, distinto en cada tubo. Mediante la acción de la enzima ADN polimerasa se incorporan diferentes nucleótidos hasta que la elongación de

la hebra complementaria se detenga al incorporar un ddNTP que no cuenta en el extremo 3' con un OH que permita continuar la síntesis de la cadena. Posteriormente los diversos fragmentos se visualizan en un gel de poliacrilamida determinando la secuencia de la hebra. Los secuenciadores de esta generación generaban lecturas con una longitud menor de 1 kb [54].

Las tecnologías de segunda generación, emplean dos mecanismos conocidos como secuenciación por ligación y secuenciación por síntesis. En la primera, una secuencia sonda que tiene en el extremo 5' un fluoróforo, se hibrida a un fragmento de ADN, el fluoróforo será captado por el medidor de fluorescencia, posteriormente se elimina el fluoróforo y sucesivas rondas de hibridación se llevan a cabo, esta tecnología fue empleada por la plataforma SOLiD, que generaba lecturas con una longitud de 50 pb [50]. La secuenciación por síntesis es empleada en los secuenciadores Illumina, donde cada nucleótido que es incorporado en la cadena que se está sintetizando emite una señal fluorescente que es captada por el sistema de lectura del secuenciador, este proceso se realiza simultáneamente en múltiples clústers [9]. Entre las principales ventajas de esta tecnología se destaca la reducción en el costo por GB secuenciada, la baja tasa de error y el alto rendimiento en la producción de datos, sin embargo la longitud de la lecturas es solo de 150 pb. En cuanto a las tecnologías de tercera generación, éstas se caracterizan por la secuenciación en tiempo real de una sola molécula de ADN, generando lecturas largas de aproximadamente 20 kb, sin requerir procesos de amplificación previos al secuenciamiento[50]. La principal limitación de estas tecnologías es la elevada tasa de error del 10 %. A continuación se presentan con mayor detalle las tecnologías empleadas para la producción de datos de la presente investigación.

2.2.1. Tecnología Illumina

La plataforma Illumina se basa en la secuenciación química por síntesis (SBS, del inglés sequencing by synthesis) que permite detectar las bases individuales a medida que éstas se incorporan en el ADN de la cadena molde, a través de desoxirribonucleótidos trifosfato (dNTP) terminadores reversibles [9]. Illumina HiSeq ofrece un alto rendimiento debido a que en una celda de flujo se inmovilizan establemente decenas de millones de moléculas de ADN molde por centímetro cuadrado. El flujo de trabajo de esta tecnología inicia con la construcción de librerías, donde el ADN es fragmentado y luego ligado a adaptadores. Posteriormente los fragmentos de ADN son fijados individualmente a una matriz sólida donde ocurren múltiples ciclos de amplificación de estos fragmentos para formar los clúster. Finalmente se da el proceso de secuenciación de los clúster incorporando dNTPs terminadores reversibles, cada uno marcado con un fluoróforo específico para cada tipo de base. Cuando uno de estos dNTPs es incorporado, su fluoróforo es liberado emitiendo una señal que es captada a través de microscopía de fluorescencia de reflexión interna total [50].

2.2.2. Tecnología GemCode -10x genomics

GemCode, es una tecnología para la preparación de librerías, que permite transformar la capacidad de los secuenciadores de lecturas cortas (secuenciadores Illumina HiSeq) para la reconstrucción de grandes fragmentos [71]. Para ello, el equipo cuenta con un chip basado en microfluidos, que permite dividir las moléculas de ADN de tal manera que todos los fragmentos producidos dentro de una partición tienen un código de barras común, permitiendo bioinformáticamente la reconstrucción de moléculas de longitudes medias entre 30 y 100 kb [46].

Para el uso de esta plataforma se debe realizar una extracción de ADN genómico de alto peso molecular, con una longitud media superior a 50 kb, teniendo como rango aceptable de concentración entre 0,8-0,2 ng/ul. Inicialmente, se tienen una serie de perlas de gel que se mezclan con las moléculas de ADN de alto peso molecular, las cuales han sido tratadas previamente con una mezcla maestra y agentes desnaturalizantes. A través de la tecnología de microfluidos se generan particiones o "gotas" conocidas como GEMs (del inglés Gel Bead-In-Emulsion), donde cada GEM contendrá una perla de gel y una sola molécula de ADN en promedio. Posteriormente, dentro de cada GEM se disuelve la perla de gel y se liberan cebadores que contienen un adaptador compatible con el cebador Illumina (secuencia R1) y un código de barras de 10X genomics de 16 pb, y cebadores con una secuencia aleatoria de 6 pb. Las GEM con los cebadores son sometidos a un proceso de incubación isotérmica donde se genera la amplificación de los fragmentos de ADN y de esta manera todos los fragmentos amplificados que provienen de la misma molécula de ADN quedan marcados con el mismo código de barras. Esto permite su posterior ensamblaje en contigs. Finalmente, hay una fase de limpieza donde se remueve el aceite de la emulsión y las librerías quedan listas para el secuenciamiento por la plataforma Illumina [46].

2.2.3. Tecnología Pacific Biosciences - PacBio

La plataforma PacBio ha sido clasificada como una tecnología de secuenciación de tercera generación, debido a que el proceso de secuenciamiento ocurre en tiempo real y no requiere una pausa entre los pasos de lectura [92], esto se debe a que el secuenciador emplea una celda especializada conocida como guía de onda de modo cero (ZMW, zero-mode waveguides, por su sigla en inglés). El ADN a secuenciar es una molécula bicatenaria, la cual se circulariza al ligar adaptadores de horquilla en cada extremo, generando una molécula de ADN circular cerrado monocatenario, conocido como SMRTbell. El proceso de secuenciamiento ocurre en la parte inferior de cada ZMW, donde se encuentra inmovilizada la enzima polimerasa, la cual se une al adaptador de la molécula diana. Los procesos de replicación ocurren en todas las ZMW cuando la polimerasa escinde el fluoróforo unido a un desoxirribonucleótidos trifosfato (dNTP), siendo captado por un sensor [50]. Dependiendo del tiempo de vida de

la enzima polimerasa, la molécula diana puede ser secuenciada múltiples veces generando lecturas largas continuas (CLR, del inglés continuous long reads), con información continua de los adaptadores y las hebras. Estas lecturas pueden ser cortadas de acuerdo a los adaptadores generando sub-lecturas conocidas como secuencias circulares consenso (CCS, del inglés circular consensus sequences).

2.3. Ensamblaje genómico

Tradicionalmente el ensamblaje de secuencias se ha definido como la reconstrucción computacional de una secuencia larga a partir de múltiples lecturas de corta longitud [39], para ello grupos de lecturas se agrupan en contigs y los contigs en scaffolds. Los contigs proporcionan una alineación múltiple de las lecturas de las secuencias, más la secuencia consenso, y a través de los scaffolds se obtiene el orden, orientación y el tamaño de los gaps entre los contigs [80]. No obstante esta definición debe ser reevaluada debido a la generación de lecturas largas por las tecnologías de tercera generación, que aportan significamente en el llenado de gaps, unión de scaffolds y solución de regiones repetitivas [92].

Existen dos tipos de enfoque de ensamblaje, *de novo* y comparativo o por referencia. El primero se refiere a la reconstrucción de una secuencia contigua sin hacer uso de una secuencia de referencia, mientras que el ensamble comparativo emplea un genoma de referencia como guía [39].

2.3.1. Algoritmos de ensamblaje

El ensamblaje de secuencias se basa en algoritmos de grafos. Un grafo es la representación de un conjunto de objetos a través de nodos o vértices que están conectados por aristas o ejes [80]. Los diferentes caminos que se pueden formar entre los nodos se conocen como camino Hamiltoniano (cada nodo es visitado solo una vez) y camino Euleriano (cada arista es visitada solo una vez) [37]. Los algoritmos para ensamblaje principamente usados son: grafos de Bruijn y Overlap-Layout-Consensus (OLC).

- **Grafo de Bruijn**

Los grafos de Bruijn se construyen a partir de subsecuencias de las lecturas con un tamaño k, conocido como k-mer. Cada subsecuencia es representada por un nodo y los ejes se establecen a través de una relación de sobrelapamiento de k-1. El objetivo posterior a la construcción del grafo es encontrar el camino Euleriano a partir del cual se construye la secuencia consenso [24]. SOAPdenovo2, Velvet, ABySS y Euler son algunos ensambladores que basan su

funcionamiento en el algoritmo de grafo de Bruijn [37].

En el caso de SOAPdenovo2 [72], éste se compone de seis módulos: correción de errores, construcción del grafo Bruijn, ensamblaje de contigs, alineamiento de lecturas pareadas, construcción de scaffold, y cerrado de gaps. El primer módulo consta de cuatro etapas obligatorias y una opcional. En la primera etapa se realiza una estimación de un espectro de frecuencias de k-mers, para esto se estiman dos tipos de k-mers, consecutivos y con espacio. A partir de esta estimación se construye una tabla de frecuencias de k-mers, que es empleada en la segunda etapa para dividir la información en dos categorías, baja frecuencia de k-mers y alta frecuencia de k-mers. Los k-mers estimados con baja frecuencia se considerarán como candidatos a errores y pasan a la tercera etapa donde ocurre una corrección de bases, empleando un algoritmo de votación rápido llamado "FAST", donde se itera sobre todas las bases posibles para sustituir el error, luego verifica la autenticidad de los nuevos k-mers generados que corresponden al k-mer donde se encontraba la base con error. La cuarta etapa de la fase de corrección emplea el algoritmo "DEEP", que tiene como objetivo realizar la corrección de las bases adyacentes o cercanas, así como los errores en los bordes de las lecturas que no se corrigieron en la etapa anterior, finalizada esta corrección se busca la subcadena más larga de la lectura en la cual los k-mers no tienen errores. Las lecturas que permanecen con errores son descartadas.

En el segundo módulo se realiza la construcción del grafo de Bruijn empleando las lecturas corregidas. En el tercer módulo se realiza el ensamblaje de contigs a través de una estrategia multi-kmer, donde el software evalúa el uso de los k-mers de tamaño pequeño para distinguir entre errores de secuenciamiento y fusionar regiones altamente heterocigotas, y k-mers de tamaño largo para resolver pequeñas repeticiones. El cuarto módulo realiza el alineamiento de las lecturas pareadas de acuerdo al tamaño de inserto permitiendo usar la información de éstas en el quinto módulo donde se realiza la construcción de scaffolds. Finalmente, se lleva a cabo la fase de cerrado de gaps, haciendo uso de la información de los alineamientos, donde cada base debe estar soportada por 80 % de las lecturas para realizar la extensión entre scaffolds.

- **Grafo overlap layout consensus**

El algoritmo overlap-layout-consensus se fundamenta en encontrar el camino Hamiltoniano, originalmente fue empleado para el ensamble de datos provenientes del secuenciamiento por Sanger, pero posteriormente fue optimizado para el ensamblaje de genomas grandes y actualmente es empleado en los ensambladores de Celera, Arachne, CAP y PCAP. Estos ensambladores realizan un pre-cálculo a través de todas las lecturas para seleccionar los candidatos de solapamiento, para ello emplea los k-mer identificados como semillas de alineación,

y construye un grafo de superposición a través del cual se obtiene la secuencia consenso [80].

Actualmente el algoritmo overlap-layout-consensus es empleado en versiones modificadas para el ensamblaje con lecturas largas provenientes de tecnologías de secuenciación de tercera generación como Pacbio y Oxford Nanopore. El software Canu [67] implementa este algoritmo, denominado jerárquico donde numerosas etapas de sobrelapamiento y corrección se llevan a cabo entre las lecturas largas, con el objetivo de mejorar la calidad de la lecturas antes del ensamblaje.

Canu se compone de tres fases, corrección, recorte (o trimming) y ensamblaje [67]. Inicialmente las lecturas empleadas se procesan a través de la estrategia conocida como MinHash overlapping, basada en un algoritmo probabilístico para detectar todos los posibles sobrelapamientos entre lecturas largas. Para hacer esto, el algoritmo estima el índice de similitud de Jaccard, comprimiendo las secuencias o lecturas en una serie de huellas representativas (o fingerprints) que consisten en un conjunto de k-mers mínimos (o minimum-valued k-mers) o k-mers con el mínimo valor. Dado que existen repetidos k-mers en el genoma, a éstos se les asigna menor peso debido a que son poco informativos sobre el origen de la lectura, en contraste los k-mers con única copia son más informativos y tendrán un mayor peso. De acuerdo con el anterior criterio se implementa una ponderación conocida como el *tf-idf* weight (term frecuency, inverse document frecuency), donde el peso asignado a cada k-mer es la combinación multiplicativa del número de ocurrencias de un k-mer en una lectura o secuencia (documento) y la rareza general del k-mer entre todas las lecturas o secuencias (el cuerpo), por consiguiente la rareza de un k-mer es útil para identificar documentos similares.

En la etapa de corrección, Canu emplea toda la información de los diversos sobrelapamientos para corregir las lecturas individuales, previamente Canu implementa dos filtros para identificar cuál sobrelape debe ser empleado para corregir una lectura [67]. El primer filtro es un filtro global donde cada lectura puede ser usada como evidencia de corrección de otras lecturas, y el segundo es un filtro local donde para cada lectura se acepta o se rechaza la evidencia proporcionada por otras lecturas, permitiendo una corrección jerárquica de las lecturas. Adicionalmente Canu estima las longitudes de lectura corregidas. En la etapa de recorte o trimming se remueven las bases que no fueron soportadas por otras lecturas, junto con la eliminación de horquillas (o hairpins) de los adaptadores y secuencias quiméricas. Cada lectura queda recortada pero tratando de conservar la mayor longitud posible que sea soportada por otras lecturas a un nivel de cubrimiento y tasa de error específicos.

La fase de ensamblaje inicia con una última fase de corrección donde cada lectura es corregida según la evidencia que proveen los alineamientos superpuestos con el fin de distinguir entre diferencias reales en las secuencias (por ejemplo, polimorfismos verdaderos) de errores de secuenciación [67]. En esta etapa, las secuencias de las lecturas con errores de secuencia-

ción no son cambiadas, en su lugar se ajusta la tasa de error. Posterior a la última fase de corrección, Canu realiza la construcción del mejor grafo de sobrelapes (o BOG del inglés Best Overlap Graph), considerando las mejores superposiciones. A partir del BOG, se construyen los contigs iniciales y para cada uno de ellos se define un perfil de error con base en las tasa de error de los sobrelapes que se usaron en su contrucción [67]. Finalmente se genera una secuencia consenso para cada contig, para ello Canu construye una secuencia molde para cada contig por medio del empalme de las lecturas que se encuentran en posiciones cercanas y que provienen de la mejor superposición.

2.3.2. Métricas de evaluación del ensamblaje

Posterior a la fase de ensamblaje, diferentes estadísticas son empleadas con el objetivo de evaluar la integridad [41] y la contigüidad del genoma ensamblado [111]. Entre las principales métricas se encuentran: tamaño del ensamblaje, número y tamaño de scaffolds y contigs, N50, L50, NG50, LG50, porcentaje de gaps, cobertura porcentual y porcentaje de mapeo de lecturas iniciales al ensamblaje (Fig **2-1**).

El N50 es un estadístico que describe la contigüidad del ensamblaje del genoma a nivel de scaffolds y contigs, asumiendo que entre más largo sea éste, mejor será el ensamblaje [111, 41]. De otra manera se puede asumir que el N50 describe un tipo de mediana de las longitudes de las secuencia ensambladas, dando mayor peso a las secuencias largas. Se calcula ordenando por longitud de mayor a menor cada scaffolds o contig. Luego, iniciando por el más largo, se suman las longitudes hasta obtener la mitad de la longitud total de todos los contigs o scaffolds en el ensamblaje [39]. No obstante, esta estadística debe ser interpretada con precaución debido a que un ensamblaje deficiente, donde se forzaron lecturas y contigs no relacionados en scaffolds, puede tener un valor grande erróneo de N50 [111]. Una estrategia para este caso es el mapeo de lecturas y las comparaciones sinténicas a gran escala para la detección de regiones erróneamente ensambladas o scaffolds quiméricos.

Una medida relacionada al N50 es el L50, el cual es empleado para representar el número de contigs y scaffolds, que son más más largos o iguales que la longitud N50 y, por ende, incluyen la mitad de las bases del ensamblaje (https://www.ncbi.nlm.nih.gov/assembly/help/). El NG50 y el LG50 son medidas con las mismas características que las métricas de N50 y el L50, pero se estiman sobre el tamaño del genoma, y no con el tamaño del ensamblaje. Estas métricas son usadas para hacer comparaciones entre ensamblajes de genomas de diferente tamaño [11].

Otra importante métrica es el tamaño del ensamblaje, el cual permite estimar la relación entre el total de bases ensambladas y el tamaño del genoma estimado con datos experimen-

tales independientes o con enfoques basados en la frecuencia de k-mers [39, 41]. Esta relación también se denomina porcentaje de cobertura que indica el porcentaje del genoma que está contenido en el ensamblaje en relación con las estimaciones de su tamaño; un rango entre 90 y 95 % es considerado como un porcentaje de cobertura bueno [111]. Sin embargo, debe considerarse que los genomas presentan un porcentaje alto de regiones repetitivas que pueden afectar el porcentaje de cobertura obtenido. También se puede considerar el porcentaje de genes que están contenidos en el ensamblaje. Adicionalmente, se estima el número total de scaffolds y contigs en el ensamblaje. El ideal es un número pequeño de scaffolds donde esté contenido el genoma. Finalmente, el porcentaje de brechas o vacíos, que son las regiones no secuenciadas entre pares de contigs o scaffolds, denominados gaps son representados como series de 'N's.

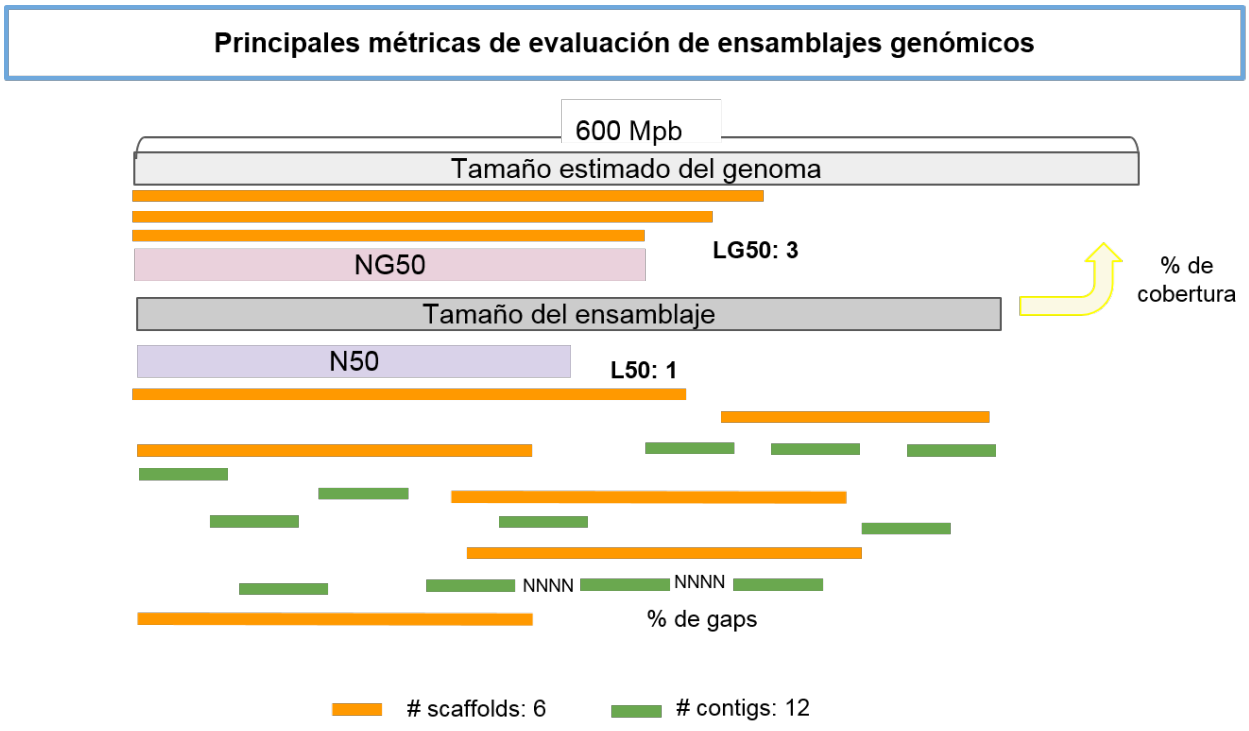


Figura 2-1.: Métricas de evaluación de ensamblajes genómicos.
En el presente ejemplo se asume un tamaño estimado del genoma de 600 Mb representado en el rectángulo de la parte inicial con color gris, este valor se toma como referencia, para estimar el estadístico NG50 (color rosado), indicando que el 50 % del ensamblaje está contenido en contigs o scaffolds iguales o mayores a este valor, el LG50 reporta el número de contigs o scaffolds que cumplen con el criterio NG50. En color gris oscuro se observa el tamaño del ensamblaje, con el cual se calcula el porcentaje de cobertura en relación al tamaño estimado del genoma. El N50 se estima de acuerdo al tamaño del ensamblaje. Los diferentes fragmentos en los que está contenido el ensamblaje a nivel de scaffolds se representan en color naranja y los contigs en color verde.

2.4. Transcriptómica

Un enfoque para el estudio de la expresión del genoma es a través del transcriptoma, el cual se asume como el conjunto de moléculas de ARN derivadas de la expresión de los genes de una célula en un momento dado y bajo un conjunto de condiciones específicas [15]; por consiguiente, el transcriptoma puede cambiar dinámicamente de acuerdo con los genes que se estén expresando [37]. Por lo anterior, la estrategia de usar diferentes tejidos facilitaria su reconstrucción.

Un acercamiento investigativo para identificar la presencia de transcritos y cuantificar el nivel de expresión de genes es a través de la hibridación de ácidos nucléicos en microarreglos, sin embargo, la capacidad para documentar y cuantificar en una forma masiva la expresión de genes en diversas condiciones experimentales por medio de esta tecnología es limitada [89]. Frente a la anterior problemática, los avances en las tecnologías de secuenciación masiva y paralela de ADN, permitieron el secuenciamiento del ARN a través del ADN complementario (ADNc) (RNA-seq) [112] y la determinación de la estructura de los genes, la identificación de sitios de inicio y finalización de la traducción, de patrones de corte y empalme alternativos (o splicing alternativo), de modificaciones post-transcripcionales, la cuantificación de los cambios en los niveles de expresión de los genes y el catálogo de ARN no codificantes [107]. Para Wang et al. (2010) una de las ventajas del uso de RNA-seq es la fácil detección de transcritos de organismos no modelo, sin genoma de referencia, adicionalmente pueden ser usados para apoyar la anotación estructural y funcional de genes [111].

La técnica de RNA-seq se basa en la secuenciación de una librería de ADNc, que ha sido obtenida del ARN mensajero presente en la muestra de ARN total [107]. A los múltiples fragmentos de ADNc se le ligan en los extremos unos adaptadores para su secuenciamiento y la obtención de lecturas cortas entre 75-400 pb según la plataforma de secuenciamiento. En los últimos años con el desarrollo de las tecnologías de tercera generación se presenta una nueva alternativa de secuenciamiento del ARN. En el caso de Pacbio, cada librería es de una sola molécula de transcripción construida a partir de ADNc, de ésta se generan lecturas con un tamaño superior a 1Kb lo que permite la identificación de los extremos 5' y 3' de las regiones no traducidas (UTR) y de los límites exón-exón que discriminan las isoformas. Sin embargo, si la secuenciación se limita a los ARN mensajeros, se obtiene un rango reducido de la expresión génica, pero con un alto potencial de precisión para la anotación [77].

La aplicación de la tecnología de RNA-seq requiere un adecuado diseño experimental que debe ser acorde al organismo de estudio y al objetivo de investigación [21]. Generalmente se pueden distinguir cuatro fases en un estudio de RNA-seq (fig **2-2**). La primera fase consiste en el diseño experimental, en la segunda fase se lleva a cabo la generación de los datos, los cuales son empleados en la tercera fase de procesamiento *in silico* y finalmente, en la cuarta

fase se desarrollan estudios sobre expresión diferencial o caracterización estructural y funcional de los diferentes genes. A continuación se detallan cada una de las fases anteriormente mencionadas.

En cuanto a la primera fase es importante destacar que el diseño experimental debe estar enfocado a una pregunta biológica, la cual orienta la fase experimental, que comprende las etapas de propagación del material biológico, y el trabajo de laboratorio donde se considera la reducción de contaminantes, para la obtención de ARN total de alta calidad. Simultáneamente se debe evaluar la estrategia de secuenciamiento, considerando el tipo de librería, profundidad de secuenciamiento y el número de réplicas. Para la construcción de la librería, el ARN total ha sido extraído previamente y el ARNm es capturado a través de la cola poli A [108], obteniéndose entre 1–2 % del ARN inicial. En cuanto al tipo de lecturas generadas, éstas pueden ser de un solo extremo usadas generalmente para estudios de niveles de expresión con especies altamente anotadas, mientras que las lecturas largas y pareadas son más informativas para caracterizar especies escasamente estudiadas [21].

Generados los datos, se inicia la tercera fase de análisis *in silico*, la cual se compone de una fase de preprocesamiento que evalúa la calidad de las lecturas, la presencia de contaminanes y adaptadores. Los datos limpios son empleados en la fase de ensamblaje del transcriptoma. Para el ensamblaje se distinguen dos enfoques: ensamblaje *de novo* y ensamblaje guiado por genoma de referencia.

El ensamblaje *de novo* se realiza cuando no se tiene genoma de referencia, o su calidad es baja. Este enfoque emplea la construcción de grafos de Bruijn a partir de k-mers provenientes de las lecturas del secuenciamiento, cada nodo es un k-mer de la lectura y los sobrelapamientos las aristas [78]. Para el ensamblaje se cuenta con diferentes ensambladores como SOAPdenovo-trans, ABySS y Trinity.

El ensamblaje guiado por genoma de referencia consiste en tres etapas, en la primera etapa se alinean las lecturas al genoma de referencia, posteriormente las lecturas sobrelapadas en cada locus son empleadas en la construcción de un grafo que representa todas las posibles isoformas, en la última etapa se atraviesa el grafo para construir todas las probables isoformas. El método anteriormente descrito es empleado en programas como Cufflinks y Trinity [83].

Trinity se compone de tres módulos: Inchworm, Chrysalis y Butterfly [48]. El primer módulo toma las lecturas de entrada, las descompone en un conjunto de k-mers sobrelapantes y para cada k-mer registra su secuencia y abundancia en la tabla hash. Posteriormente, se realiza la remoción de posibles k-mers con errores. El tercer paso es la selección del k-mer con la mayor abundancia, el cual es empleando como semilla para la reconstrucción de contigs pre-

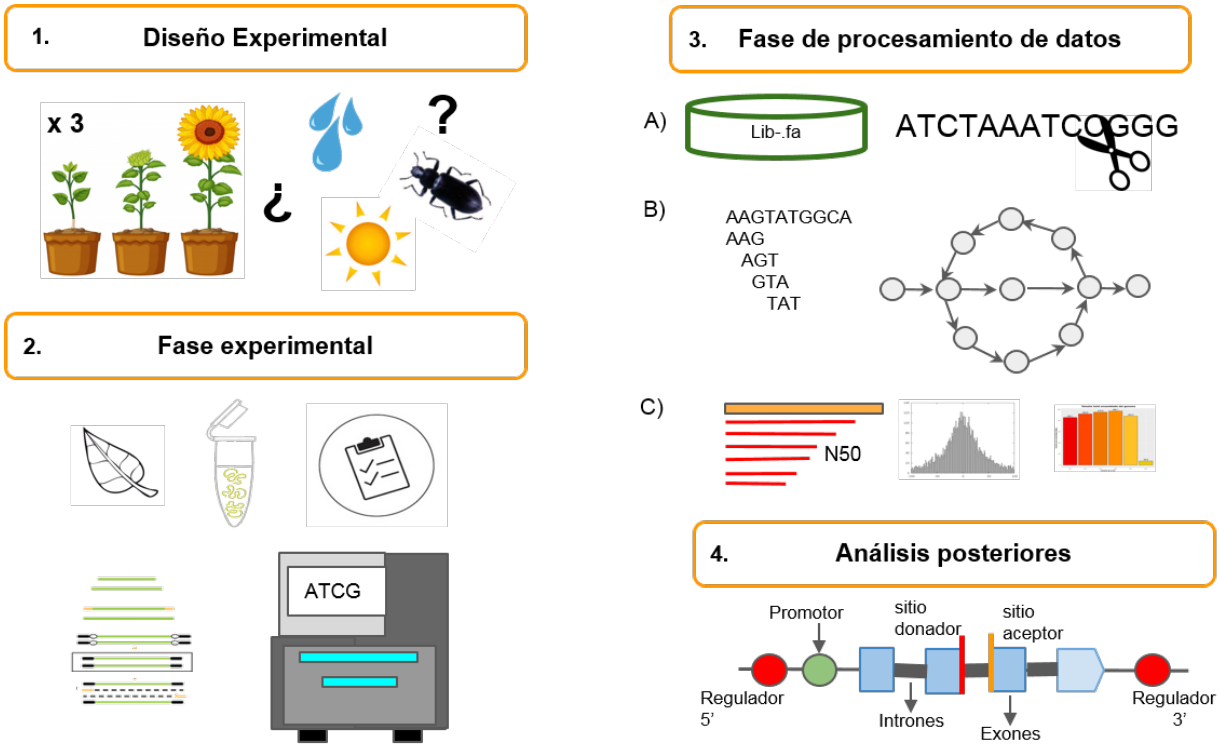


Figura 2-2.: Pipeline de análisis de datos de RNA-seq.
Las investigaciones con datos de RNA-seq, consideran cuatro etapas: **1.**Diseño experimental de acuerdo al objetivo de investigación. **2.**Fase experimental, donde se lleva a cabo el proceso de extracción del ARN, construcción de la librería y secuenciamiento. **3.**Fase de procesamiento, se realiza el control de calidad de las lecturas, limpieza de adaptadores y remoción de lecturas de baja calidad. Posteriormente se lleva a cabo el ensamblaje y evaluación del mismo. **4.**Análisis posteriores involucran el uso de los transcritos para realizar procesos de anotación del genoma o análisis de expresión diferencial.

liminares. A partir del k-mer semilla, la extensión se hace tanto en dirección 5'-3' como en dirección 3'-5'. Los k-mers sobrelapantes (k-1) que se usan para realizar la extensión a partir del k-mer semilla se van seleccionando de acuerdo a su abundancia. La extensión, base por base, se realiza hasta agotar todos los k-mers que puedan proveer una extensión. Cuando se hayan reconstruido los contigs a partir del k-mer más abundante, todos los k-mers que integran los contigs son removidos de la tabla hash. El proceso de selección de k-mers semilla y su extensión bidireccional para reconstruir nuevos contigs se realiza de manera iterativa hasta que todos los k-mers de la tabla hash se han usado.

El módulo Chrysalis emplea los contigs lineales que reporta Inchworn, a través de tres fases [48]. En la primera se realizan grupos si hay una superposición de K -1 entre ellos y si existe un número de lecturas que soportan la unión de los contigs evaluados. En la segunda fase se construye un grafo de Bruijn para cada grupo de manera individual. Posteriormente, cada lectura es mapeada al grupo con el cual comparte el mayor número de k-1 mers y el número

de lecturas mapeadas a cada grupo proporciona una medida de soporte para el grupo. Los grupos con muy bajo soporte son descartados.

El módulo Butterfly se compone de dos fases [48]. En la primera fase (conocida como trimming) se remueven posibles errores de secuenciación al identificar las aristas de los nodos que son soportadas por relativamente pocas lecturas. En la segunda fase (conocida como graph compaction) se realiza una reconciliación entre los nodos de los grafos de Bruijn individuales generados por Chrysalis para producir grafos más compactos (con secuencias más largas). Las fases 1 y 2 se repiten de manera iterativa hasta alcanzar convergencia. Este módulo usa un procedimiento de programación dinámica que identifica los caminos (o paths) a través de los grafos de Bruijn que están mejor soportados por las lecturas y pares de lecturas (paired-ends) reales y que representan posibles transcritos, sus isoformas y copias parálogas.

Finalizada la fase de reconstrucción del transcriptoma, cada ensamblaje debe ser evaluado. Sin embargo, las métricas empleadas en el ensamblaje genómico como N50 y NG50 no son adecuadas para seleccionar el mejor ensamblaje del transcriptoma debido a que las secuencias más largas o el ensamblaje total más grande pueden indicar un nivel de sobreinclusión o quimerismo [106]. Una alternativa para esta problemática es considerar un conjunto de métricas primarias (salidas básicas de un ensamblador) y las métricas de los ensamblajes post-procesamiento (representa un ensamblaje más refinado). Entre las evaluaciones primarias se encuentran: número de secuencias ensambladas, longitud media, tasa de alineación, el número de transcritos ensamblados que cubren completamente un transcrito de referencia y una alta recuperación de ortólogos ultra conservados [57]. De acuerdo con la evaluación primaria se obtendrá un transcriptoma con una calidad superior, que puede ser empleado en etapas posteriores para predicción funcional y estructural de los genes a través de bases de datos como UniProtKB, SwissProt o Pfam [111] y análisis de expresión diferencial [112].

2.5. Anotación genómica

La anotación genómica es un paso fundamental en la caracterización estructural y funcional de un genoma, centrando su atención en regiones que corresponden a elementos de relevancia biológica, tales como genes codificantes de proteínas, ARN no codificantes o regiones repetitivas [41]. Este proceso comprende dos fases: anotación estructural y anotación funcional. La fase de anotación estructural tiene como objetivo la identificación de la ubicación y la estructura de todos los genes de un genoma, incluyendo las regiones no traducidas (UTR) y las isoformas de un gen [65], mientras que la anotación funcional permite relacionar las anotaciones estructurales con su producto (proteína o ARN) y la función asignada al producto génico, tarea que puede ser llevada a cabo con el uso de base de datos especializadas para búsqueda de dominios funcionales, ortólogos relacionados, redes metabólicas y ontología de

genes, entre otros [111].

La anotación estructural puede abordarse desde un enfoque comparativo (basado en homología) o no comparativo (basado en evidencia). El primero emplea la información de especies cercanas altamente caracterizadas para predecir secuencias codificantes de genes a través de la similaridad en la secuencias de proteínas. De otra manera el enfoque no comparativo emplea datos de RNA-seq, EST, cDNA, para identificar y caracterizar los genes. En el caso de los genes no expresados o no caracterizados se puede realizar predicciones *ab initio* basadas en homología.

En la anotación desde un enfoque basado en evidencia, la identificación de genes con base en datos de RNA-Seq, puede abordarse desde tres enfoques: el primero emplea un genoma de referencia al cual se alinean las lecturas en cada locus, reconstruyendo los transcritos. En el segundo, se realiza un ensamblaje *de novo* de las lecturas, que posteriormente se alinean contra un genoma de interés para deducir posibles estructuras de genes. Finalmente, el tercer enfoque integra los datos de RNA-Seq, ya sean ensamblajes o lecturas, en programas de predicción de genes codificadores de proteínas [65].

La anotación desde un enfoque de homología asume la similitud de secuencias a nivel de proteínas y dominios conservados, para realizar alineamientos múltiples entre especies que pueden estar alejadas evolutivamente. Bajo esta misma perspectiva de homología pero en genomas de especies evolutivamente relacionadas, se considera la posibilidad de mapear coordenadas de inicio y final de los exones de un genoma fuente a un genoma diana, esta estrategía es conocida como anotación por mapeo.

Al igual que en el ensamblaje genómico, en la fase de anotación es necesario realizar su evaluación, para ello se considera intersectar el número de familias de proteínas compartidas con la mayoría de los eucariotas, calculando el porcentaje de grupos ortólogos universales de copia única. Para llevar a cabo el proceso de anotación han surgido diversas pipeline como Maker y PASA, que buscan la integración de las predicciones *de novo* con los datos experimentales. Maker está basado en una arquitectura modular, integrando diferentes programas como Repeatmasker, BLAST, Exonerate, Augustus, SNAP, entre otros. Adicionalmente amplía las clases de Bioperl, GenericHit y GenericHSP para facilitar el análisis comparativo y la anotación automática[18].

El proceso de anotación con Maker ocurre en cinco pasos **2-3**: calcular, filtrar/agrupar, pulir (o polish), sintetizar y anotar. El primer paso tiene como objetivo la búsqueda de regiones repetitivas mediante RepeatMasker y la identificación regiones de EST, mRNA y proteínas con significativa similaridad con el genoma evaluado mediante BLAST. En el segundo paso de filtrado y agrupamiento, se realiza la identificación y eliminación de predicciones y

alineamientos que no cumplen con una serie de criterios previamente establecidos. Con el agrupamiento se identifican los sobrelapes entre predicciones y alineamientos, permitiendo la agrupación de datos y la identificación de evidencia redundante. El tercer paso de pulido usa Exonerate a partir de los hits de BLAST, para obtener mayor precisión entre los límites de exones. En el cuarto paso Maker sintetiza toda la evidencia que tiene, la cual es empleada por SNAP, para modificar el modelo de Markov y realizar predicciones *ab initio*, adicionalmente se pueden realizar predicciones a través de otros programas como Augustus. Finalmente, en el quinto paso se integran las predicciones de SNAP o del programa empleado con las evidencias y se genera la anotación [18, 17].

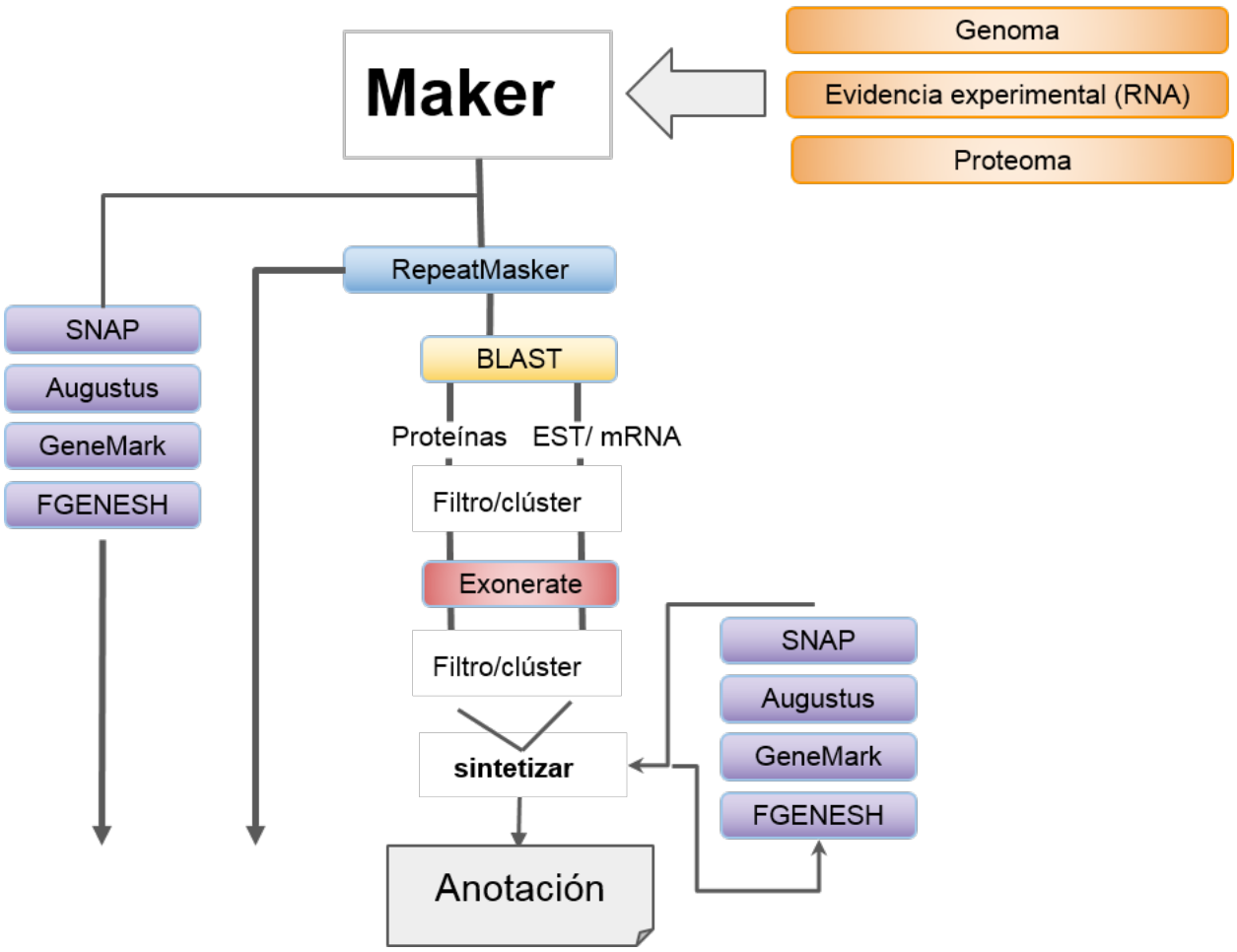


Figura 2-3.: Pipeline de anotación - maker

La pipeline de Maker inicia con predicciones *de novo* (color morado), simultáneamente se realiza la identificación de regiones repetitivas (color azul). Posteriormente se emplean BLAST,para encontrar regiones similares. En rojo se presenta la fase de polish, seguida de la fase de síntesis generando la anotación.

2.6. Genómica comparativa

La genómica comparativa es un enfoque poderoso que permite acercarse a comprender la dinámica evolutiva de los genes y de los genomas [113], y permite comprender la información contenida en el ADN de múltiples genomas a través de la búsqueda de regiones conservadas ricas en genes ortólogos, es decir genes derivados de un antecesor común [91]. Los avances en las tecnologías de secuenciamiento han permitido obtener ensamblajes genómicos de alta calidad con su respectiva anotación, haciendo posible análisis comparativos de genes ortólogos en diferentes especies. No obstante, la identificación de genes ortólogos ha sido un reto debido a la evolución dinámica de los genomas y la acumulación de cambios en éstos [102].

Para inferir genes ortólogos se han empleado métodos basados en la similitud de las secuencias y su reconstrucción filogenética [3]; el primero realiza la comparación de genomas y agrupa los genes por alta similaridad para identificar grupos de ortólogos, a través del algoritmo de clusters de Markov. El segundo usa familias candidatas de genes determinadas por similaridad, y luego con base en un análisis donde se fusionan las filogenias de genes y de especies se determinan subconjuntos de ortólogos a través de algoritmo de agrupamiento jerárquico de genes correlacionados, guiado por las relaciones filogenéticas [102]. Otra perspectiva es a través de la localización de loci y la conservación sinténica en el orden y la orientación de los genes a lo largo de los cromosomas [102]. De acuerdo a las anteriores líneas de trabajo, las herramientas computacionales se han basado en la construcción de árboles desde un enfoque filogenético y la elaboración de gráficos 2D donde cada eje representa una secuencia contigua en el genoma [60].

Para Tang et al. (2011) la principal dificultad biológica de la que se debe partir es la detección de verdaderos ortólogos y su distinción de los parálogos, dicha problemática se incrementa cuando se analizan genomas que han pasado por eventos de duplicación completa del genoma, por lo tanto la construcción de bloques sinténicos pueden ser 1:1 , 1:2 , 2:2, entre otras dependiendo de cuándo ocurrió el evento de duplicación en la historia evolutiva de ambos linajes. El análisis comparativo en el reino vegetal es un reto debido a los eventos de duplicación completa del genoma, reordenamientos genéticos y translocaciones génicas [113]. En cuanto a los métodos que definen los bloques sinténicos a través del encadenamiento de genes homólogos putativos, asignando una puntuación (score) a cada bloque y aplicando un criterio de corte, generando representaciones en gráficas de puntos, donde cada punto representa un par de marcadores de genes putativos en secuencias similares y los bloques sinténicos se representan como diagonales. Sin embargo, algunos métodos no diferencian la edad evolutiva de los bloques ni la identidad de los bloques

2.7. Historia evolutiva del frijol Lima

Las plantas cultivadas que sostienen la creciente población humana se originaron mediante el proceso de la domesticación hace unos 10.000 años en varios continentes [53]. Durante este proceso, las plantas se adaptaron al ambiente de cultivo creado por el hombre mediante respuestas fisiológicas y cambios morfológicos que en su conjunto se conocen como el síndrome adaptativo de la domesticación y que incluyen cambios en el hábito de crecimiento, en el tamaño de la planta y partes de la planta (por ejemplo de la semilla y el fruto), en la pérdida del mecanismo de dispersión natural de la semilla, entre otros. De estos rasgos, la pérdida del mecanismo de dispersión de la semilla es uno de los más importantes, no sólo porque permitió la adaptación de las plantas silvestres al ambiente de cultivo sino que la dispersión de semillas es un rasgo indeseable en los sistemas de producción debido a que reduce el rendimiento de los cultivos. Por lo tanto, entener el control genético de este rasgo es de interés en cultivos de semillas como las leguminosas.

Uno de los cultivos más importantes de leguminosas es el frijol, especialmente del género *Phaseolus*. El género *Phaseolus*, de origen americano, cuenta con más de 50 especies silvestres pero sólo se han reportado cinco especies domesticadas, a saber, *Phaseolus vulgaris* L. (el fríjol común), *Phaseolus coccineus* L., *Phaseolus dumosus* MacFac., *Phaseolus acutifolius* A. Gray y *Phaseolus lunatus* L. (el fríjol Lima). Estas cinco especies pertenecen filogenéticamente a dos linajes separados, el linaje vulgaris que incluye a las primeras cuatro especies domesticadas mencionadas y otras especies silvestres principalmente de Mesoamérica, y el linaje lunatus que incluye a *P. lunatus* y otras especies silvestres principalmente de Suramérica [33].

El proceso de domesticación en el frijol común y en el frijol Lima ocurrió a través de múltiples e independientes eventos [25, 100, 4], que generaron una serie de cambios morfológicos y fisiológicos donde se destacan el aumento del tamaño de la semilla, cambio de hábito de crecimiento (de indeterminado a determinado), pérdida de la dormancia de la semilla y cambios en el mecanismo de dispersión de la semilla [40, 47]. Estos cambios ocurrieron en las poblaciones silvestres fundadoras favoreciendo la adaptación al medio ambiente de cultivo de las variedades criollas [25], las cuales han sido clasificadas en los acervos mesoamericano y andino, teniendo como criterio el tamaño de la semilla y la distribución geográfica.

En el caso del fríjol común, éste explora un rango altitudinal entre los 1.200 y 2.500 metros sobre el nivel del mar (msnm) para el acervo Andino, caracterizado por un tamaño de semilla grande [16], mientras que el acervo mesoamericano comprende el rango de 1.500-1.900 msnm con semilla pequeña [82]. En el fríjol Lima se estableció para el estado silvestre [95], que el acervo Mesoamericano se caracteriza por presentar semilla pequeña, su distribución comprende México, Centroamérica, el Caribe, Colombia y la vertiente oriental de los Andes de

Perú, Bolivia y norte de Argentina. Su rango altitudinal varía entre los cero y 1.600 msnm, tolerando incluso climas extremadamente secos de la costa peruana (para los materiales cultivados) y bosque de transición entre las regiones muy húmedas y secas (semideciduos y Caducifoliate) con temperaturas superior a 2°C [8]. Mientras que la semilla del acervo Andino se caracteriza por su gran tamaño y su distribución comprende desde los 400 hasta los 2.000 msnm en la vertiente occidental de los Andes de Ecuador y norte del Perú, llegando en ocasiones a climas templados de altitudes superiores a 2.500 msnm, y un clima tropical húmedo de la región amazónica entre Perú y Ecuador [8]. De acuerdo a lo anterior, se puede concluir que el fríjol Lima explora un rango altitudinal más amplio que el frijol común, el cual comprende desde cero hasta 2.500 msnm, tolerando diferentes grados de temperatura y explorando diferentes ambientes [95, 8] mientras que el fríjol común está reducido al rango de 1.200 a 2.500 msmn, y por ende a los ambientes ecológicos que se ubican en dicho rango.

En cuanto a los anteriores dos acervos genéticos, investigaciones con la región ITS (internal transcribed spacer, por su sigla en inglés) del ADN ribosomal [100] y polimorfismos del ADN nuclear y del cloroplasto [4] establecieron tres acervos génicos en el fríjol Lima silvestre denominados Mesoamericano I (MI), Mesoamericano II (MII) y Andino (A). El origen Andino del fríjol Lima había sido validado por Fofana et al. (1999) mediante análisis filogenéticos de polimorfismos del ADN del cloroplasto. Para la clasificación de las poblaciones domesticadas del fríjol Lima, también se emplea el criterio de la forma y tamaño de la semilla, encontrándose semillas pequeñas y redondeadas (cultivares Sieva y Potato) para las razas mesoamericanas y semillas grandes y aplanadas (cultivares Big Lima) para la razas andinas.

El fríjol común es la leguminosa de grano más importante debido a su alto contenido nutricional y distribución global de siembra [88] que supera los 23 millones de toneladas al año [12]. En segundo lugar se encuentra el fríjol Lima, que es equiparable nutricionalmente con el fríjol común [59, 56], sin embargo su producción mundial es inferior a 200 mil toneladas anuales [8], por consiguiente el consumo de frijol lima es menor que el de frijol común.

A nivel genómico, los estudios han estado enfocados en frijol común, reportándose el secuenciamiento y ensamblaje del genoma de referencia para dos genomas que representan los dos acervos (andino y mesoamericano). El genoma del acervo andino fue ensamblado en 473 Mb [98] y el del acervo mesoamericano en 549.6 Mb [105], con N50 a nivel de scaffold de 50.4Mb y 433.8 Kb, respectivamente. Adicionalmente, el frijol común cuenta con diversos ensambles de transcriptoma en diferentes tejidos, estados de desarrollo, factores bióticos y abióticos, generando un atlas de expresión [110, 7, 62, 105], así como también investigaciones comparativas sobre la conservación del genoma entre especies modelo de leguminosas [19, 96, 98].

Para frijol Lima a la fecha no existía genoma de referencia, solo se contaba con el reporte del ensamblaje de un transcriptoma desde un enfoque de estrés biótico específico empleando

el hongo *Trichoderma viride*, con el objetivo de identificar genes relacionados con la divergencia ambiental y posibles adaptaciones en el género *Phaseolus* [69]. A nível de genómica comparativa no existen investigaciones para el frijol Lima. Sin embargo, una aproximación a la organización estructural de esta especie, fue realizada a través de mapas citogenéticos, reportando reordenamientos estructurales en los cromosomas 2, 9 y 10 de frijol Lima y un alto nivel de macro-colinearidad en los cromosomas 3, 4 y 7 con respecto al frijol común. [10, 6].

Objetivos

General:

- Generar una estrategia computacional para la detección de microsintenia en regiones genómicas asociadas a genes de domesticación en frijol Lima.

Específicos:

- Ensamblar *de novo* scaffolds del genoma de frijol Lima (variedad G27455 Colombia-Sucre) a través de la combinación de macro y microfragmentos genómicos secuenciados por la tecnología Illumina.
- Ensamblar *de novo* el transcriptoma de tres tejidos vegetales de frijol Lima para apoyar la anotación de los scaffolds.
- Detectar bloques microsintenicos asociados a genes de domesticación con especies evolutivamente cercanas al género *Phaseolus* para validar la organización de estos genes entre las especies analizadas.

3. Ensamblaje genómico de novo de alta calidad de Frijol Lima

3.1. Resumen

El Frijol Lima (*Phaseolus lunatus* L.) es la segunda especie domesticada más importante del género *Phaseolus*, después de frijol común (*Phaseolus vulgaris* L.). Estas dos especies comparten similitudes a nivel taxonómico, ecológico y evolutivo. Sin embargo, las investigaciones genómicas en este género han estado enfocadas en el frijol común, dejándose especies huérfanas como ha ocurrido con el frijol Lima, una especie de gran importancia agronómica y ecológica en la que no se han desarrollado las herramientas genómicas necesarias para avanzar en el estudio de genes de interés, ni para comprender la estructura genética de este cultivo.

En la presente investigación se reporta el primer secuenciamiento y ensamblaje *de novo* del genoma de frijol Lima del genotipo G27455 del acervo Mesoamericano, a través de la combinación de lecturas provenientes de tecnologías de secuenciamiento de segunda (Illumina y 10x-Genomics) y tercera generación (PacBio). Se generaron 97.6 Gb de datos, que fueron empleados en una estrategia de ensamblaje que evaluó siete tamaños de k-mer y la reconstrucción del genoma mediante un enfoque jerárquico. El genoma de frijol Lima se ensambló en 541 Mb, contenido en 496 contigs con una contigüidad 5.5 Mb (N50).

Conclusión: Se realizó el primer ensamblaje de alta calidad del genoma de frijol Lima de una accesión colombiana, con alta continuidad y calidad por base, mediante la implementación de una estrategia de ensamblaje que empleó datos provenientes de diferentes tecnologías de secuenciamiento. Este primer ensamblaje aporta al conocimiento genómico de esta especie, generando los primeros scaffolds del genoma de frijol Lima.

Palabras claves: *Phaseolus lunatus* L., secuenciamiento genómico, ensamblaje

3.2. Introducción

La familia de las fabáceas o leguminosas agrupa a más de 20.000 especies en tres subfamilias: Mimosoideae, Caesalpinioideae y Papilionoideae [19]; en esta última se encuentran las especies de cultivo más importantes para la dieta humana, razón por la cual se han desarrollado diversas investigaciones con el objetivo de caracterizar sus genomas e identificar genes asociados a rasgos de importancia agronómica. De la amplia diversidad de esta familia, el género *Phaseolus* alberga uno de los cultivos más antiguos del mundo, el frijol.

El frijol común es la especie más importante de este género, seguido de frijol lima. Estas dos especies comparten una interesante historia evolutiva, especialmente en cuanto al proceso de domesticación ocurrido a través de eventos múltiples e independientes [25, 100, 4] donde las variedades criollas divergieron de sus parentales silvestres, siendo clasificadas en los acervos mesoamericano y andino [95, 8]. A nivel ecológico el frijol Lima explora un amplio rango altitudinal que comprende desde cero hasta 2.500 msnm, tolerando diferentes grados de temperatura y explorando diferentes ambientes [95, 8]. A nivel genómico el frijol Lima es una especie diploide con 11 pares de cromosomas (2n=2X=22) [82, 31] y tamaño estimado del genoma de 600 Mb [2]. Sin embargo, el estudio de esta especie a escala genómica no había sido posible, debido a que a la fecha no había genoma disponible.

El ensamblaje genómico de alta calidad caracterizado por su continuidad y calidad por base [79] facilita el desarrollo de una amplia gamas de investigaciones para la caracterización funcional y estructural del mismo, permitiendo la identificación de elementos de relevancia biológica como regiones repetitivas, ARNs no codificantes o genes codificantes de proteínas [41]. La identificación de estos elementos se logra mediante diferentes estrategias tales como la comparación de las secuencias de ADN (genómica comparativa) con especies estrechamente relacionadas o la incorporación de evidencia experimental a través de datos de RNA-seq [21]. En el caso de un genoma vegetal, las investigaciones han sido enfocadas en la identificación y caracterización de genes involucrados en rasgos agronómicos como rendimiento, hábito de crecimiento, floración, mecanismos de resistencia a enfermedades y respuestas específicas a estrés biótico o abiótico [36]. Otro importante aspecto de indagación ha sido la búsqueda de relaciones entre los indels y su incidencia en el proceso de domesticación, junto con la dinámica evolutiva del genoma [101]. De acuerdo a lo anterior, el potencial investigativo del genoma es amplio, aportando en múltiples aspectos al conocimiento de las especies que cuentan con este recurso.

Una reciente tendencia para el secuenciamiento *de novo* de genomas complejos es la combinación de plataformas de secuenciamiento de segunda generación (Illumina) y de tercera generación (Oxford Nanopore, Pacific Biosciences-PacBio), con el objetivo de aumentar el tamaño de los scaffolds ensamblados y corregir los errores surgidos en el secuenciamiento

[50]. En el caso de Illumina, esta tecnología se basa en la secuenciación por síntesis química (SBS) que permite detectar las bases individuales a medida que se incorporan en el ADN de la cadena molde [9], mientras que Pacbio emplea la enzima ADN polimerasa unida a nanoestructuras conocidas como "Zero Mode Waveguides" (ZMW), las cuales captan en tiempo real señales de fluorescencia a medida que se sintetiza la hebra complementaria [84, 94].

El proceso bioinformático que se usa para el ensamblaje genómico está estrechamente relacionado con el tipo de datos generados por secuenciamiento. A nivel algorítmico, el proceso de ensamblaje con lecturas cortas se realiza a través de grafos de Bruijn dirigidos que usa k-mers (subcadenas de longitud k) construidas a partir de la lecturas [91], para representar sobrelapes entre las secuencias (k-1). Luego de la construcción del grafo, se identifica el óptimo camino, conocido como camino Euleriano, el cual pasa por cada arista una sola vez con el cual se construirá el ensamblaje [99].

Generalmente la estrategia de ensamblaje *de novo* para lecturas cortas se compone de tres etapas: ensamblaje de contigs, scaffolding y llenado de gaps. En la primera etapa se obtienen contigs, a partir de la implementación de un algoritmo de ensamblaje. En la segunda etapa los contigs ensamblados se conectan ordenadamente formando scaffolds mediante la superposición de contigs y el uso de datos de una segunda fuente de información como lo son las lecturas largas o lecturas marcadas con barcodes como las producidas por las tecnología 10x. Finalmente, en la etapa de llenado de gaps, los espacios son estimados a partir del tamaño de inserto de las lecturas o mediante un enfoque comparativo con una especie evolutivamente cercana. Las dos últimas etapas pueden ser realizadas iterativamente, con el objetivo de mejorar la calidad del ensamblaje [99].

El ensamblaje con lecturas largas emplea algoritmos conocidos como Overlap-Layout-Consensus (OLC). Inicialmente el algoritmo OLC realiza un pre-cálculo a través de todas las lecturas para seleccionar los candidatos de solapamiento, empleando k-mers identificados como semillas de alineación entre los pares de lecturas, a partir de éstos se construye un grafo de superposición donde se busca el camino Hamiltoniano (cada nodo es visitado sólo una vez) [37] y al recorrer el camino con los mejores sobrelapes se obtiene la secuencia consenso [80].

En la estrategia de ensamblaje *de novo* para lecturas largas se distinguen tres etapas: corrección, limpieza y ensamblaje [67]. Inicialmente las lecturas proporcionadas al software ensamblador como Canu, implementan una estrategia conocida como MinHash overlapping para la identificación de regiones similares entre pares de lecturas. En la etapa de corrección se emplea toda la información de los diversos sobrelapamientos para corregir las lecturas individuales. En la segunda etapa se remueven las secuencias que no fueron soportadas por otras lecturas, junto con la eliminación de adaptadores. En la tercera etapa se realiza la construcción del grafo de sobrelapes, considerando las mejores superposiciones, las cuales

generan una secuencia consenso para cada contig [67]. En la figura **3-1** se observan las diferentes etapas de ensamblaje de acuerdo al tipo de lectura empleada.

En la presente investigación se generó el primer ensamblaje de alta calidad del genoma de frijol Lima, de un cultivar colombiano. Este genoma contribuye al conocimiento de la familia de las fabáceas, un grupo de alto interés agroeconómico puesto que en éste se encuentran las legumbres de cultivo más importantes para el consumo humano [45, 88].

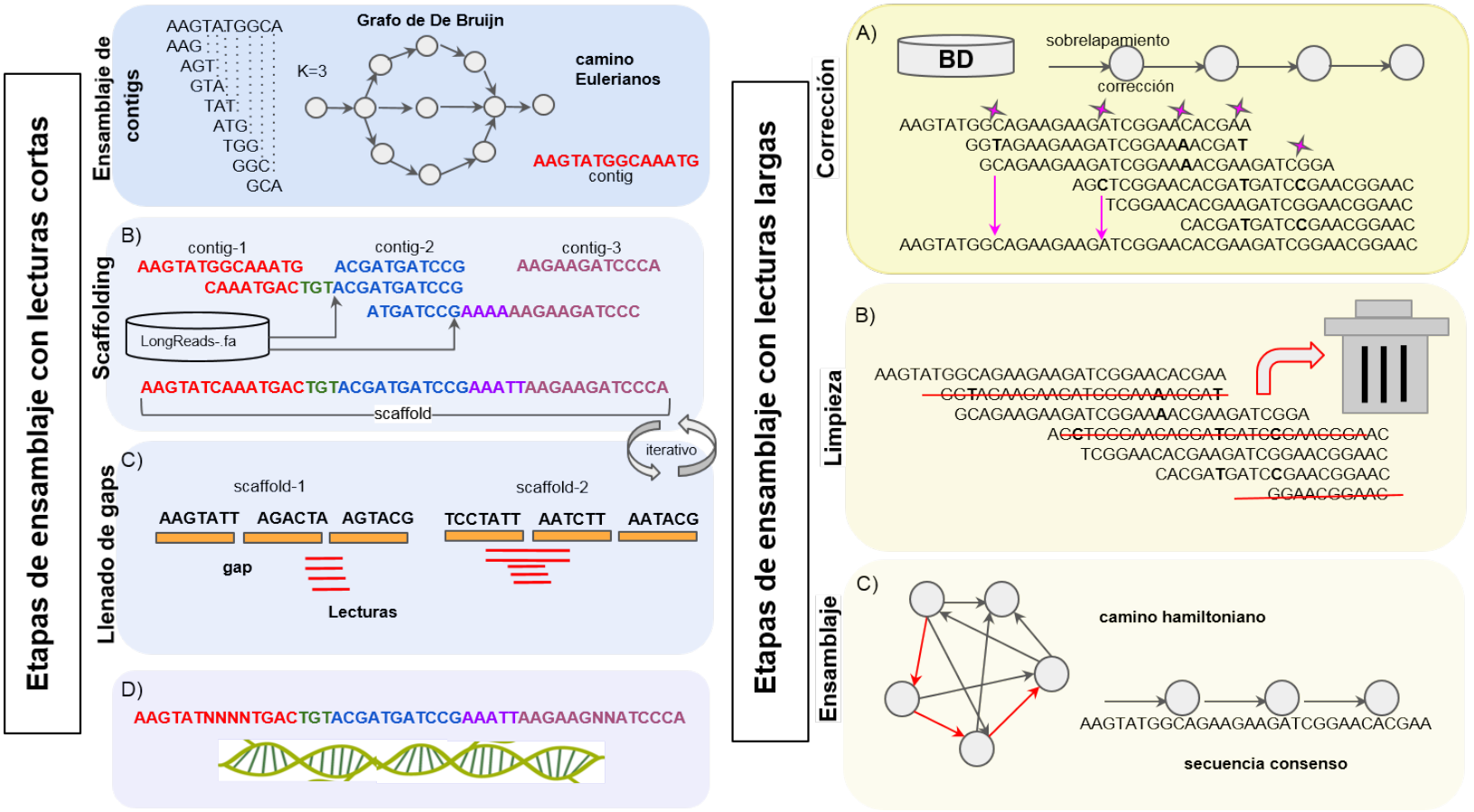


Figura 3-1.: Aproximación metodológica para el ensamblaje genómico a partir de lecturas cortas y lecturas largas.
En el ensamblaje genómico basado en lecturas cortas, se distingue tres etapas principales: **A)** Ensamblaje de contigs, donde las lecturas se dividen de acuerdo a un tamaño de subsecuencia k, conocido como k-mers. A partir de estos k-mers se construye el grafo de Bruijn, el recorrido del camino Euleriano del grafo genera el ensamblaje de los contigs. **B)** Scaffolding, permite la unión de contigs. **C)** Llenado de gaps, usa la información de las lectura pareadas para resolver regiones con secuencias 'N'. En el ensamblaje con lecturas largas desde un enfoque jerárquico se diferencian tres etapas: **A)** Corrección, las lecturas se sobrelapan buscando la corrección de errores. **B)** Limpieza, las lecturas sin soporte identificadas en la fase anterior se eliminan. **C)**Ensamblaje, se realiza empleando el grafo Overlap-Layout-Consensus, buscando el camino Hamiltoniano que reconstruye la secuencia consenso.

3.3. Resultados y Discusión

3.3.1. Conjunto de datos de secuenciamiento

Para la presente investigación se generaron 243 millones de lecturas que contienen 61.9 giga bases (Gb), a través de la combinación de tecnologías de secuenciamiento Illumina y Pacbio. Para el secuenciamiento con la plataforma Illumina se obtuvieron 36.3 Gb en dos librerías de lecturas pareadas, una libreria con tamaño de inserto de 450 bp, y la segunda librería fué construida mediante la tecnología 10x-Genomics con tamaño de insertos entre 500 y 700 bp. Con la tecnología Pacbio se generaron 25.6 Gb en una librería construida a partir de moléculas mayores a 3kb. En total se originaron 97.6 Gb de datos crudos de secuenciamiento. Se obtuvó un cubrimiento de 103.17 X, suponiendo un tamaño de 600 Mb del genoma de frijol Lima. En la tabla **3-1** se especifican los rendimientos obtenidos por cada plataforma de secuenciamiento.

Tabla 3-1.: Producción de datos de secuenciamiento genómico obtenidos por la tecnología Illumina y PacBio.

Libreria	Plataforma	Longitud promedio de lectura (pb)	Número de lecturas (M)	Bases totales(Gb)	Datos crudos (Gb)
1	Illumina	150	103	15.5	31
2	Illumina 10x-Genomics	150	138	20.8	41
3	Pacbio	11914	2	25.6	25.6
		Total	243	61.9	97.6

[1] pb: pares de bases, M: millones, Gb: gigabase.

3.3.2. Evaluación de la calidad de las lecturas

Los 103 millones de lecturas (31 Gb) de la librería 1, presentaron un 35 % de contenido GC, y calidad en el llamado de las bases superior a 28 Q (anexo A, fig:**A-1**). Aunque la calidad de las lecturas fue alta, el análisis de contenido de secuencias sobrerrepresentadas (Anexo A, fig:**A-2**) evidenció la necesidad de realizar un corte de las primeras 15 bases en el extremo 5' y de cinco bases en el extremo 3', por lo cual se removió el 0.1 % de las lecturas totales.

En cuanto a la librería construida con la tecnología 10X, la longitud de las lecturas disminuyó a 120 pb, debido a los barcodes presentes en el extremo 5' de cada lectura que son empleados en esta tecnología.(Anexo A, fig:**A-3** y fig:**a-4**).

3.3.3. Estrategia de ensamblaje de novo del genoma de Frijol Lima

El ensamblaje *de novo* del genoma de frijol Lima se llevó a cabo bajo cuatro enfoques: el primero consistió en el ensamblaje de las lecturas provenientes de la tecnología Illumina (librería 1) con la evaluación de seis tamaños de k-mers (51, 61, 71, 81, 91 y 101). El segundo enfoque incorporó las lecturas de la tecnología de 10X-Genomics (librería 2) a la librería 1 con evaluación de dos tamaños de k-mers (K-71 y K-81). El tercer enfoque empleó un kmer de tamaño 48, empleando solamente las lecturas provenientes de la librería dos. Finalmente, en el cuarto enfoque se utilizaron en un único ensamblaje las lecturas provenientes de la tecnología de PacBio. La figura **3-2** resume las estrategias de ensamblaje que se usaron, donde el color naranja y azul representan ensamblajes mediante grafos de Bruijn y el color rojo el ensamblaje por OLC. En total se obtuvieron diez ensamblajes *de novo*.

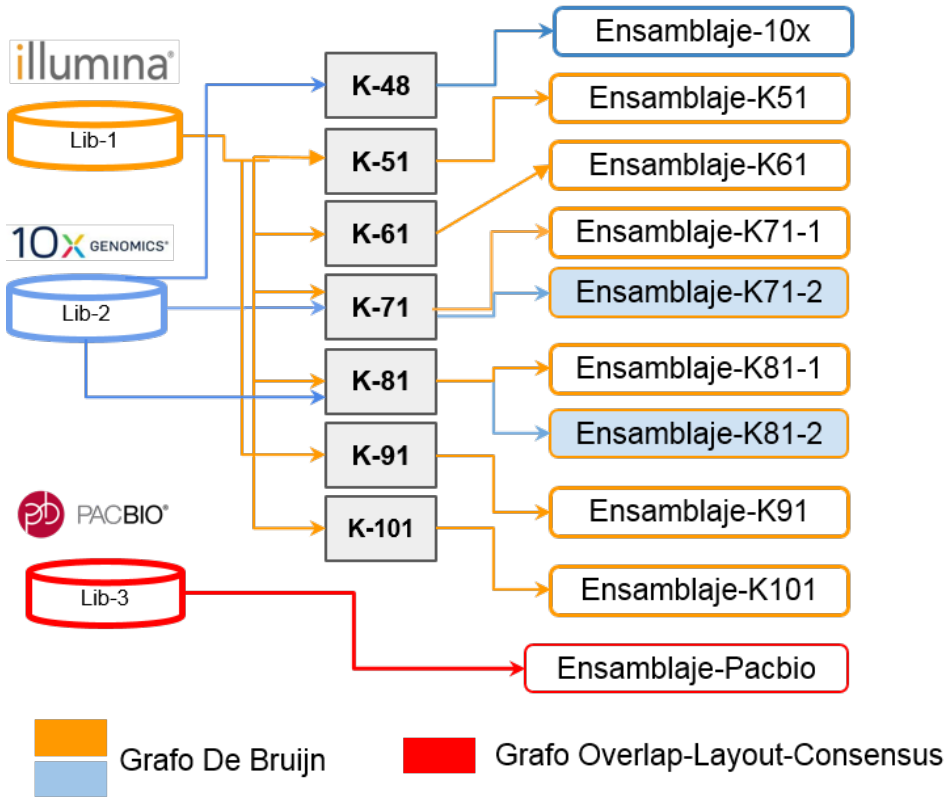


Figura 3-2.: Estrategia de ensamblaje *de novo* del genoma de frijol Lima.

La librería uno (color naranja), se empleó con seis tamaños de k-mer (51,61,71,81,91 y 101) generando un ensamble por cada k-mer. La librería dos (color azul) se ensambló mediante dos estrategias, la primera empleó la pipeline supernova que usa un tamaño de k-mer de 48, haciendo uso de los barcodes de la tecnología 10X-Genomics. La segunda estrategia integró la librería uno y dos, evaluando dos tamaños de k-mer (71 y 81), los ensamblajes de este enfoque se representan como K71-2 y K81-2. La librería tres (color rojo) se ensambló de manera independiente generando el ensamblaje-Pacbio.

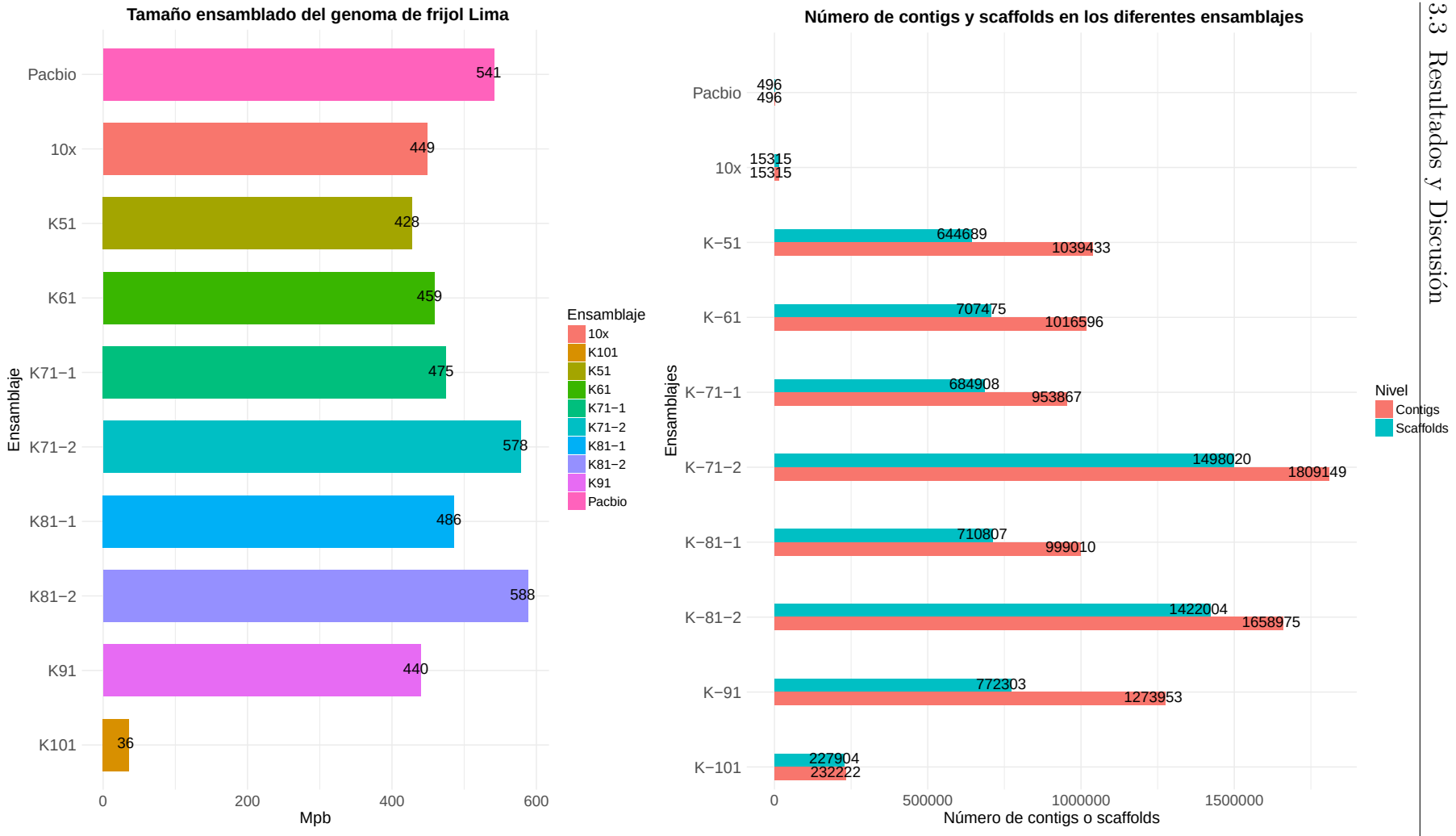
De los diez ensamblajes realizados, al considerar únicamente el tamaño ensamblado del genoma (figura **3-3** -(a)) se encuentra que el mayor tamaño fue de 588 Mb, obtenido con la combinación de las dos librerías secuenciadas con Illumina y empleando un k-mer de 81 (ensamblaje-K81-2), seguido del ensamblaje-K71-2 con 578 Mb. El ensamblaje-Pacbio tuvo un tamaño de 541 Mb. En cuanto a los ensamblajes que emplearon únicamente la librería 1, el mayor tamaño ensamblado fue con K-81 con 486 Mb, seguido de K-71-1 con 475 Mb y K-61 con 459 Mb, mientras que el tamaño más grande de k-mer evaluado de 101 tan solo logró un tamaño ensamblado de 36.8 Mb, siendo el menor entre el conjunto de k-mers evaluados. Al emplearse únicamente la libreria dos, con la tecnología de barcode de 10x, se obtuvo un ensamblaje de 449 Mb. La diferencia de los tamaños ensamblados del genoma con diferentes k-mers evidencia el posible impacto del tamaño de las subsecuencias en los que se cortan las lecturas y que son empleadas en la construcción del grafo de ensamblaje [39].

En cuanto al número de contigs y scaffolds (figura **3-3**-(b)), se observa que el genoma está contenido en el menor número de contigs en el ensamblaje-Pacbio con 496 contigs, seguido del ensamblaje con k-mer 101 con 227.904, sin embargo el tamaño ensamblado con el k-mer 101 fue el más bajo (36.8 Mb). Con respecto a los ensamblajes K81-2 y K71-2, que presentaron los mejores tamaños de ensamblaje (588 Mb y 578 Mb, respectivamente), se observa que el número de contigs y scaffolds fue el más alto entre los ensamblajes, evidenciando una baja contigüidad de los ensamblajes, y por ende una baja relación entre tamaño ensamblado y el número de contigs o scaffolds en los que está reconstruido el genoma.

Al realizar la comparación de los ensamblajes mediante el estadístico N50 (Tabla **3-2**) a nivel de contigs (figura **3-4**), el mejor valor observado fue para el ensamblaje-Pacbio con 5536 kb, seguido del ensamblaje 10x con 58 kb, y con un menor desempeño se encuentran los ensamblajes para Illumina K-101 con 0.157 kb seguido de K-91 con 0.595 kb. En cuanto al nivel de scaffolds, el valor máximo de N50 lo obtuvo el ensamblaje 10x con 2580 kb, sin embargo su magnitud es inferior a la obtenida a nivel de contig para el ensamblaje-Pacbio. El k-mer 101 mostró el menor valor de N50 para scaffolds (0.158 kb). De acuerdo al tamaño ensamblado, el número de contigs y el N50, se considera que la mejor reconstrucción del genoma de frijol Lima se obtuvo con el ensamblaje-Pacbio al lograrse un genoma de 541 Mb de un estimado de 600 Mb, representado en tan solo 496 contigs y un N50 de 5.5 Mb.

Tabla 3-2.: Estadístico N50 de los diferentes ensamblajes (contigs y scaffolds) obtenidos a partir de los datos de secuenciamiento en las plataformas Illumina (K51 hasta K101), 10X-Genomics y PacBio

	Ensamblajes									
N50	K51	K61	K71-1	K71-2	K81-1	K81-2	K91	K101	10x	PacBio
Contig (pb)	2736	3364	3624	1402	2225	2286	595	157	58480	5536000
Scaffold (pb)	19123	18064	16738	11293	12654	11316	4556	158	2580000	5536000



(a) (b)

Figura 3-3.: Métricas de los ensamblajes genómicos obtenidos a partir de los datos de secuenciamiento en las plataformas Illumina (K51 hasta K101), 10X-Genomics y PacBio.

(a) Tamaño ensamblado del genoma según la estrategia empleada de ensamblaje

(b) Número de scaffolds y contigs en cada ensamblaje

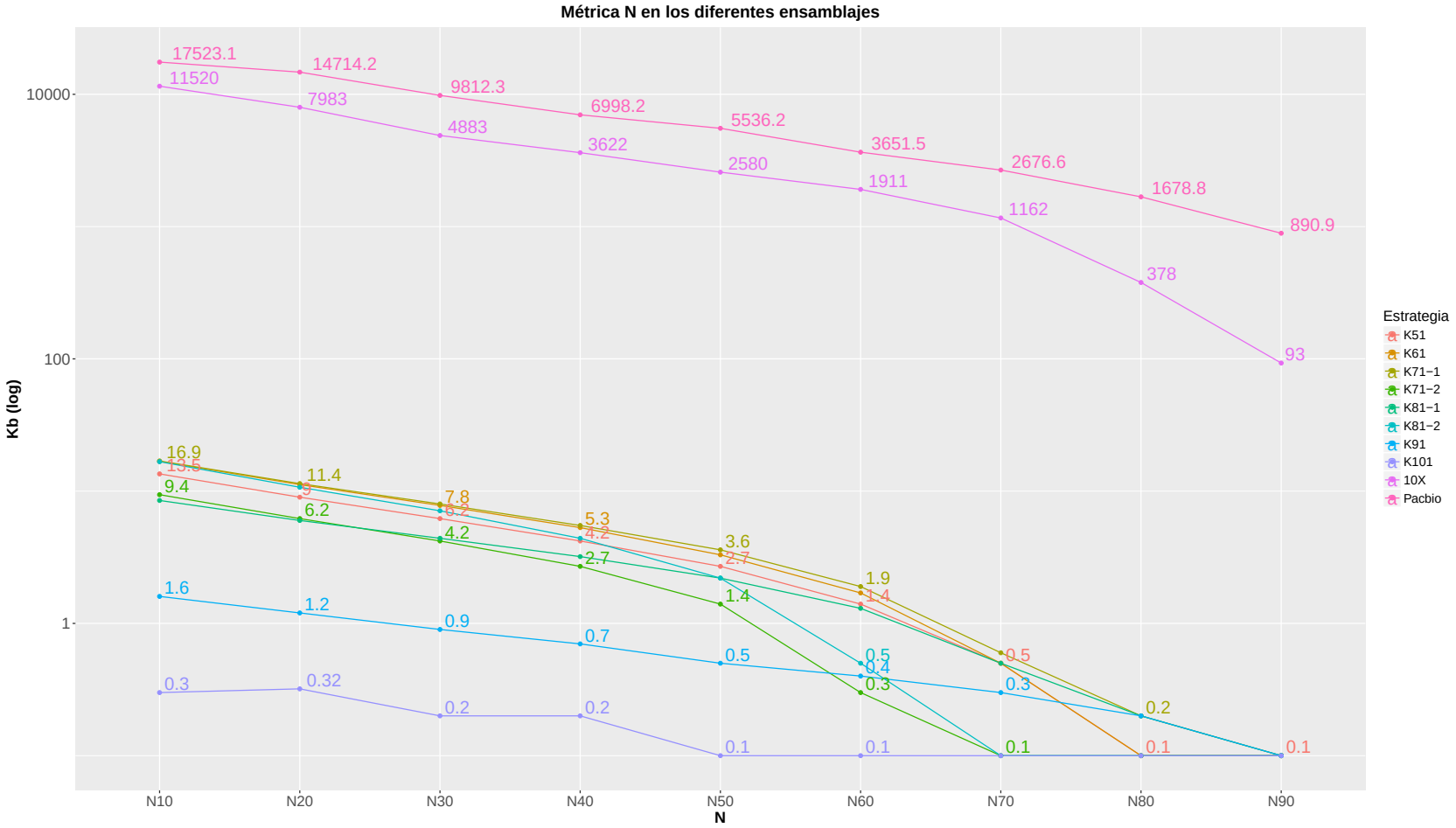


Figura 3-4.: Métricas de evaluación de ensamblajes genómicos obtenidos a partir de los datos de secuenciamiento en las plataformas Illumina (K51 hasta K101), 10X y PacBio.

- **Evaluación del contenido génico en los diferentes ensamblajes**

Para evaluar el contenido génico de los diferentes ensamblajes, se alineó un conjunto de 1440 genes ortólogos de copia única provenientes de la base de datos Embryophyta(odb9) a cada ensamblaje. En la figura **3-5** se observa que el ensamblaje Pacbio presentó la mayor cantidad de genes ortólogos con 1333 (86.4 % de copia única y 6.2 % duplicados para un total de 92.6 %), seguido del ensamblaje K81-2 con 1290 genes (89.6 %). El ensamblaje K-91 presentó el 19 % (275) de genes fragmentados, siendo el mayor valor para este criterio; en los restantes ensamblajes el porcentaje de fragmentación fue menor o igual al 7 %. Los ensamblajes con mayor porcentaje de genes perdidos fueron K-101 (100 %), K-91(46.3 %) y K81-1 (11.9 %); en los restantes ensamblajes este porcentaje fue inferior al 8 %, siendo el ensamblaje Pacbio el que presentó el menor valor con 5.9 %. De acuerdo a la evaluación del contenido génico se evidencia que el ensamblaje Pacbio presenta el mayor contenido génico de los ensamblajes evaluados.

Adicionalmente, como criterio de comparación se realizó la misma evaluación de contenido génico con el conjunto de 1440 ortólogos en las versiones 1.0 y 2.1 del genoma de frijol común. Se estimó para la versión 1.0 un contenido de ortólogos del 92 % (1324 ortólogos) y del 92.2 % (1331 ortólogos) para la versión 2.1; esta comparación evidencia la competitividad del ensamblaje Pacbio obtenido para frijol Lima con un contenido de ortólogos del 92.6 %.

En cuanto a los genes reportados como perdidos se encontraron 64 genes compartidos por todos los ensamblajes, de éstos se encontró que el 21 % pertenecen a proteinas con dominios Pentatricopeptide repeat (PPR), las cuales han sido caracterizadas como mediadoras el procesamiento, empalme, edición, estabilidad y traducción de ARN [75], el 14 % pertenecen a proteinas con dominio Tetratricopeptide repeat (TPR), su función está asociada a faciliar la interacción proteína-proteína, estando presente en varias proteínas funcionalmente diferentes, relacionadas con funcionamiento de chaperonas, ciclo celular, proceso de transcripción y los complejos de transporte de proteínas [13]. Un 10 % está relacionado con proteinas kinasas, las cuales regulan la actividad biológica de las proteínas mediante la fosforilación de aminoácidos específicos con ATP, induciendo un cambio conformacional de una forma inactiva a una activa de la proteína [38], es importante destacar que estas proteínas tienen múltiples copias en el genoma por consiguiente no es posible hallar un único ortólogo.

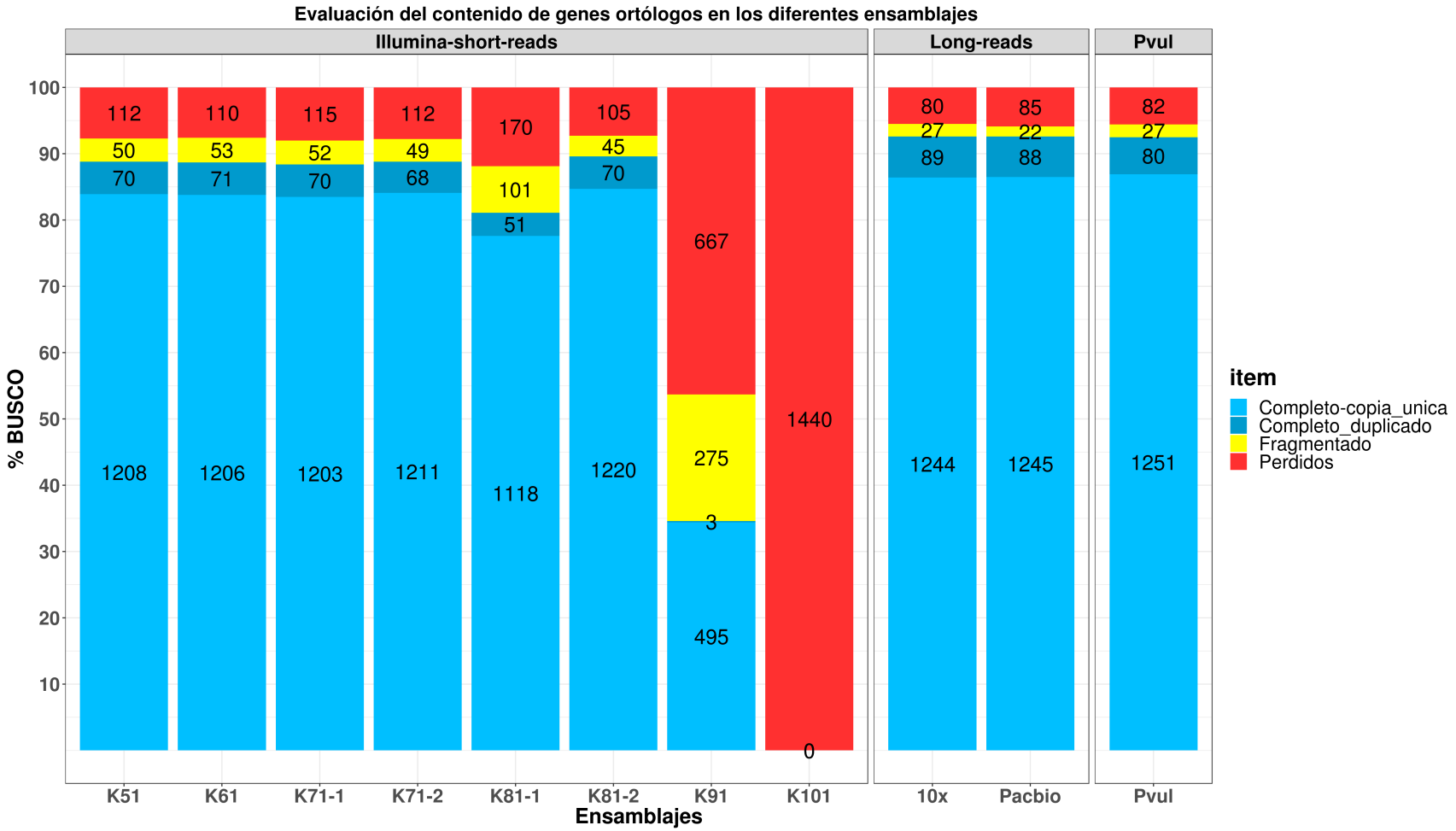


Figura 3-5.: Evaluación del contenido génico en los diferentes ensamblajes

- **Evaluación de los ensamblajes con datos de GBS**

Con el objetivo de verificar posibles contaminantes en las lecturas de las diferentes librerías secuenciadas y comparar las tasas de alineamiento de lecturas cortas en nuestros ensamblajes, frente a las generadas con el genoma de frijol común, se mapearon datos de genotipado por secuencuenciamiento (GBS) obtenidos de 10 accesiones de frijol Lima (ver Tabla **3-4** en la sección de materiales y métodos) a cada ensamblaje (fig **3-6**). Se encontró que para todas las accesiones evaluadas, las mayores tasa de mapeo se obtuvieron en el ensamblaje-Pacbio. De las muestras del acervo mesoamericano, la accesión G25787-D presentó la tasa más alta de mapeo con el 95.85 %, en contraste del 68.45 % al emplearse el genoma de frijol común para esta misma accesión. Estas 10 accessiones se utlilizaron previamente en un estudio de genómica poblacional en frijol Lima que usó como referencia el genoma del frijol común [29], lo cual significa que las mayores tasas de mapeo observadas cuando se usa el genoma de frijol Lima como referencia, permitirá aumentar el número de variantes genéticas en futuros estudios.

La tendencia observada en las accesiones del acervo mesoamericano se mantiene en las muestras del acervo andino, donde las accesiones G25108-D y G26468-W, con el 94.82 % y el 94.91 % en tasas de mapeo respectivamente, presentan los valores más altos. Es importante resaltar que las accesiones que presentaron las mayores tasas de mapeo corresponden tanto a accesiones domesticadas (G25787-D y G25108-D) como silvestres (G26468-W), lo cual evidencia una alta representación del genoma de esta especie en el ensamblaje-Pacbio. En cuanto a los restantes ensamblajes que emplearon lecturas de Illumina, el promedio de tasa de mapeo fue del 90 %, lo cual era lo esperado por ser variedades de la misma especie.

Al comparar las tasas de mapeo entre frijol común y frijol Lima se observa un incremento aproximado del 30 % para los ensamblajes con secuencias de frijol Lima, sugiriendo un alto porcentaje de pérdida de datos en las investigaciones que empleaban como genoma de referencia el frijol común con datos de frijol Lima; a pesar que estas dos especies son muy cercanas evolutivamente, se evidencia la importancia de generar un ensamblaje de alta calidad para frijol Lima, que contribuya a comprender a nivel genómico esta especie y establecer posibles relaciones con rasgos agronómicos de interés.

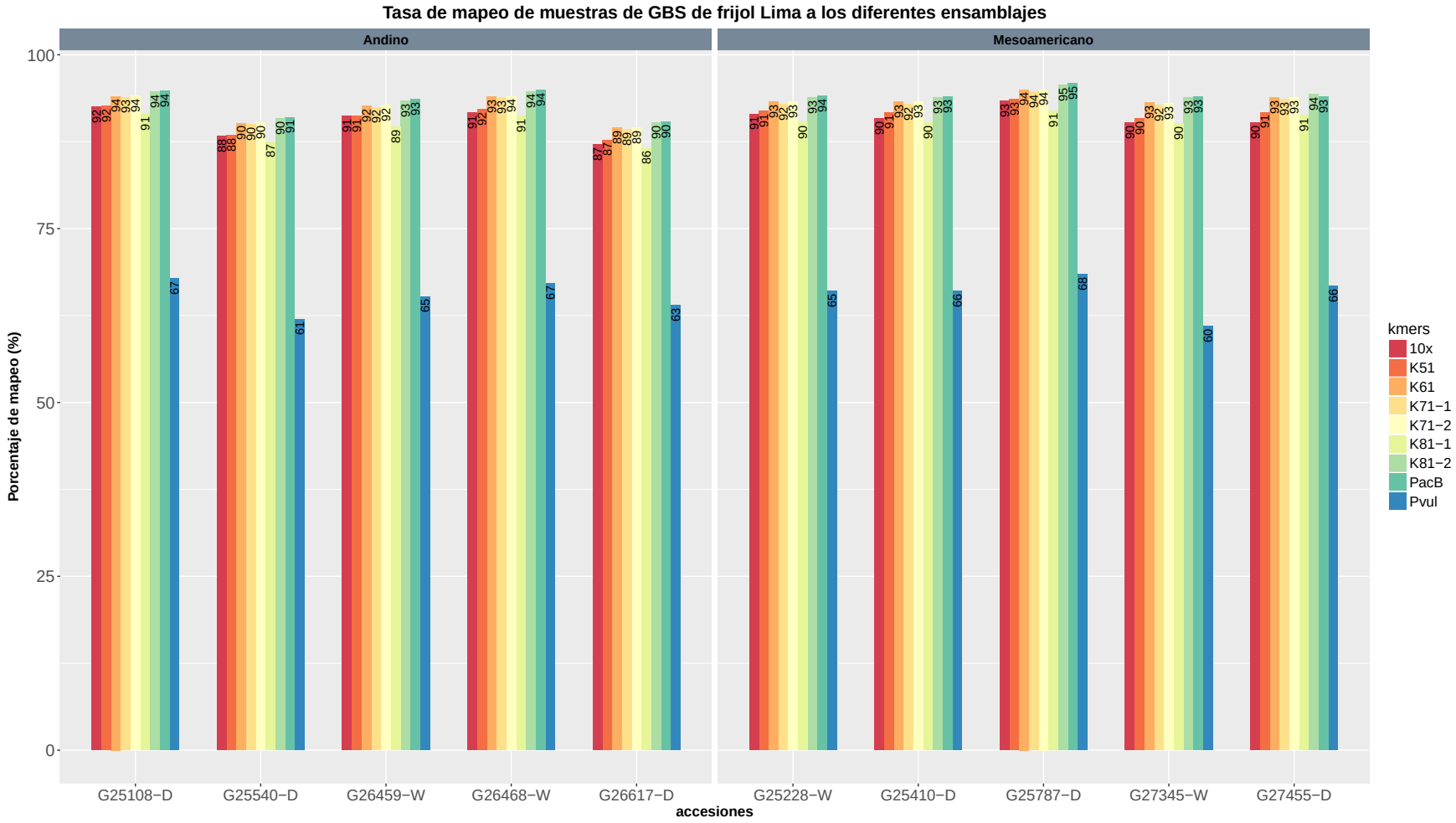
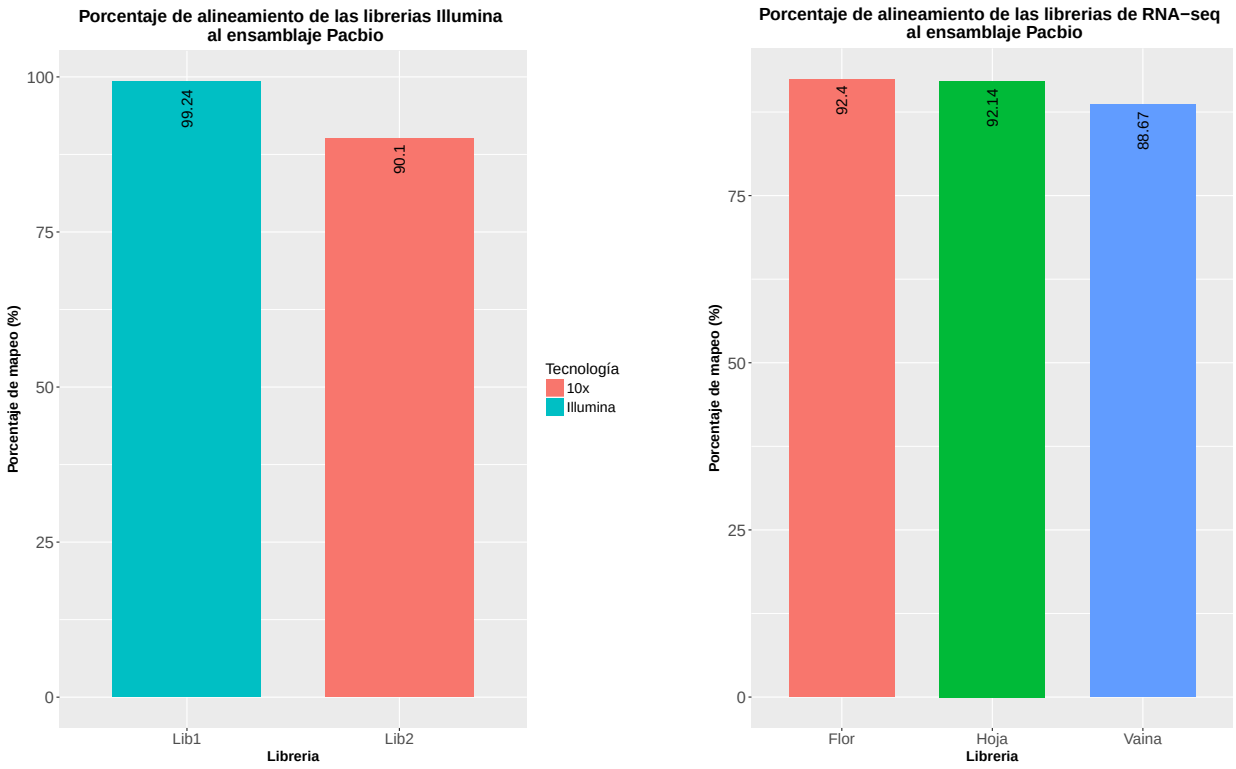


Figura 3-6.: Tasas de mapeo de lecturas obtenidas por la técnica de genotipado por secuenciación (GBS), en un conjunto de 10 accesiones silvestres (W) y domesticadas (D) de frijol Lima, a los diferentes ensamblajes del genoma de frijol Lima obtenidos por las tecnologías Illumina (K-51 a K81-2), 10X-Genomics y PacBio.

3.3.4. Validación del genoma de alta calidad de frijol Lima

Para confirmar la cobertura en el conjunto de contigs finales del ensamblaje-Pacbio, se alinearon todas las lecturas de la librería de Illumina (librería 1) y 10X (librería 2) a este ensamblaje. Las tasas de alineamiento (Fig **3-7** -(a)) fueron del 99.24 % para la librería 1 y del 90 % para la librería 2. Los resultados de la evaluación de cobertura sugieren que nuestros contigs finales cubren la mayor parte del genoma de frijol Lima.

Adicionalmente, se alinearon lecturas de secuenciamiento de RNA-seq de frijol Lima de tres librerías (hoja, flor y vaina) al ensamblaje Pacbio (Fig **3-7** -(b)). Se observó una alta representatividad de los datos de expresión en el genoma con porcentajes superiores al 90 % para los tejidos de flor y hoja y del 88 % para el tejido de vaina.



(a) (b)

Figura 3-7.: Tasas de alineamiento de datos genómicos y transcriptómicos obtenidos con la plataforma Illumina al ensamblaje del genoma de frijol Lima obtenido con la tecnología PacBio

(a) Tasa de alineameinto de las lecturas del genoma obtenidas con la plataforma Illumina (librería 1) y 10-X Genomics (librería 2)

(b) Tasa de alineamiento de las lecturas de RNA-seq

- **Estimación del cubrimiento del genoma**

Con respecto a la distribución de cubrimiento para cada librería (fig **3-8**), se observó un rango de profundidad entre 30x y 65x, con un máximo de 46x para 12.355.627 pb de la librería 1. En la librería 2 (10X Genomics), la profundidad osciló entre 8x y 13x, con 26.612.507 pb en 10X. Los anteriores resultados de acuerdo a Ekblom & Wolf (2014)[39] son un buen indicador de la representatividad de la información genómica de la especie en estudio.

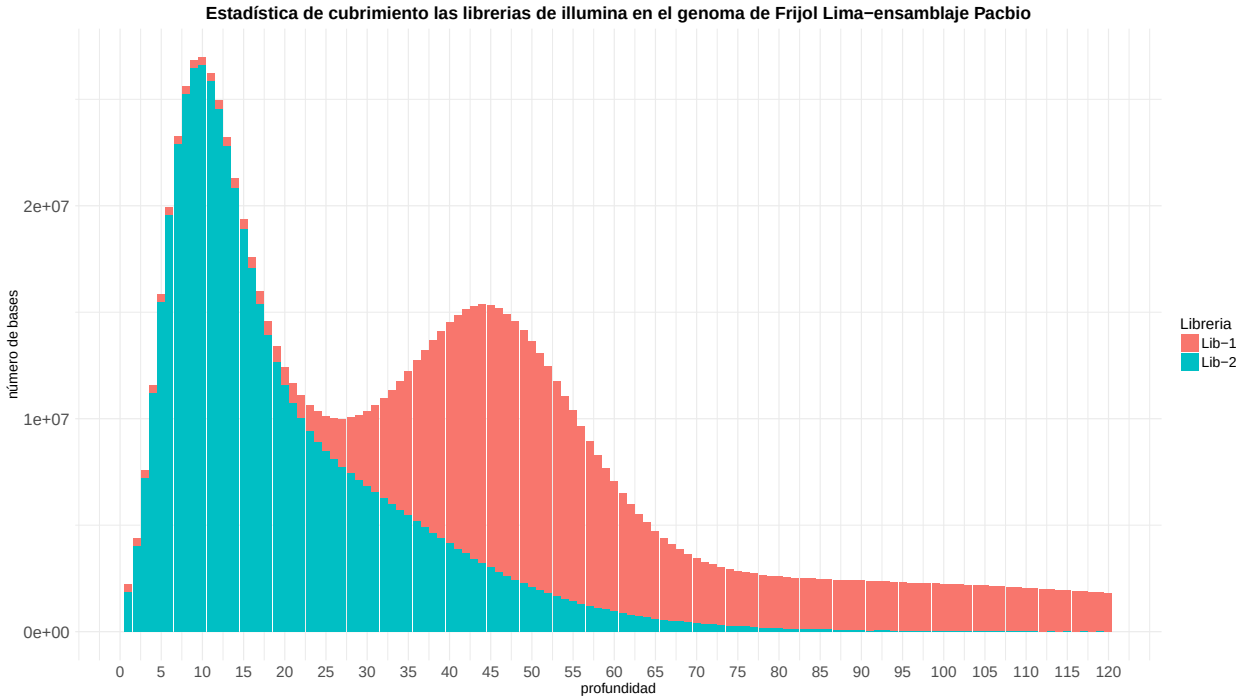


Figura 3-8.: Estadística de cobertura de la librerías Illumina (librería 1) y 10-X Genomics (librería 2) cuando son mapeadas al ensamblaje-PacBio obtenido para el genoma de frijol Lima

- **Estimación de la tasa de error**

Para la tasa de error de secuenciación, los máximos valores se encontraron en los extremos de las lecturas, en las primeras cinco bases y en las cinco últimas (Fig **3-9**). Para la librería 1, la tasa de error máxima se encontró en la base 138, a diferencia de la librería 2, que se halló en la primera posición (Tabla **3-3**). En el caso de la librería 2, esto se debe al procedimiento de construcción de la librería, donde al extremo 5' de la molécula es incorporado el barcode (16 pb) [46]. La tasa de error máxima en ambas librerías se encontró por debajo del 0.02 % y en promedio en 0.007 %, esto valores indican que la tasa de error, estuvo por debajo de un error por cada 10.000 bases, este resultado es comparable con el estándar de calidad reportado para el genoma de *Oryza sativa* Nipponbare [34], evidenciando la alta calidad del presente ensamblaje.

Tabla 3-3.: Estadísticas de tasa de error de secuenciamiento de las librerias de Illumina

Tasa de error	Lib1 (Illumina)	Lib2 (10x)
Tasa de error máxima	0.0098	0.028
Tasa de error promedio	0.0036	0.007
Base de tasa de error máxima	138	1

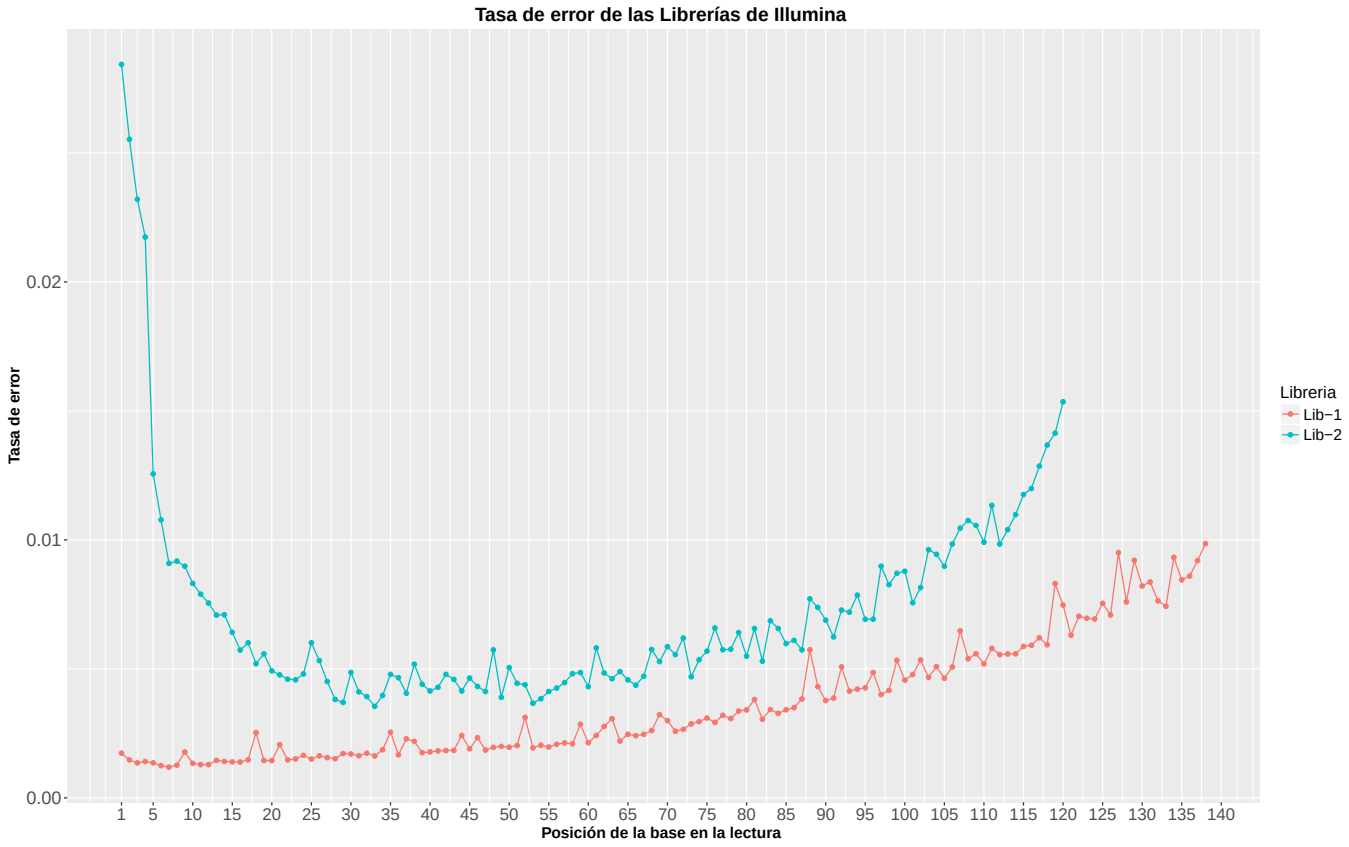


Figura 3-9.: Tasas de error del secuenciamiento de las librerias Illumina (librería 1) y 10-X Genomics (librería 2)

3.4. Conclusiones

Se generó el primer ensamblaje de alta calidad para el frijol Lima, con una significativa contiguidad en 496 contigs y un tamaño ensamblado de 541 Mb, representando el 90 % del tamaño del genoma estimado en 600 Mb. La evaluación de este ensamblaje permite suponer que es lo suficientemente completo para representar a nivel genómico esta especie, lo cual se evidencia en los mapeos de datos previos de genotipado por secuenciamiento de accesiones

silvestres y domesticadas (acervos mesoamericano y andino), y el alto contenido génico.

En cuanto a la aproximación algorítmica empleada en las estrategias de ensamblaje, se evidencia que los grafos de Bruijn, usados con lecturas cortas y un tamaño de k-mer de 71, presentó las mejores métricas de ensamblaje, frente a los restantes tamaños de k-mer evaluados. En contraste, el grafo de overlap layout consensus empleado con lecturas largas generó el mejor ensamblaje del genoma, evidenciando la diferencia que existe en contigüidad y calidad para el ensamblaje genómico entre el uso de tecnologías de tercera generación de lecturas largas y las tecnologías de segunda generación de lecturas cortas. El lograr una alta contigüidad del genoma es muy importante para el desarrollo de múltiples investigaciones que no son posibles con un genoma de baja calidad y altamente fragmentado.

3.5. Materiales y métodos

3.5.1. Obtención de ADN, construcción de librerias y secuenciamiento

El ADN genómico total fue aislado a partir de hojas jóvenes (trifolios) de la variedad domesticada G27455 (Colombia- Sucre) del acervo mesoamericano de frijol Lima. Las semillas empleadas en la propagación de las plántulas provienen del Banco de Germoplasma del Centro Internacional de Agricultura Tropical (CIAT), Cali-Colombia, las cuales fueron enviadas al proveedor del secuenciamiento.

Las datos se generaron por una combinación de tecnologías de secuenciamiento: Illumina y Pacbio. Para el secuenciamiento por la plataforma Illumina se construyó una librería con el kit NEB Next® Ultra™. En éste se emplearon entre 5 ng – 1 ug de ADN, el ADN fue fragmentado y se realizó una selección de tamaño de fragmentos, posteriormente se llevó a cabo la ligación de adaptadores. El control de calidad fue realizado en un instrumento Qubit 2.0 fluorometer (Thermo Fisher Scientific, Waltham, MA, USA) y PCR cuantitativo. Adicionalmente, la distribución de fragmentos se evaluó en un instrumento Agilent Bioanalyzer 2100 System. El promedio de tamaño de inserto fue de 450 pares de bases (pb). El secuenciamiento generó lecturas pareadas de 150 pb empleando la plataforma Illumina HiSeq (Illumina, San Diego, CA, EE. UU.),

Con la tecnología 10x se construyó una librería, se empleó 1 ug/uL de ADN de acuerdo al protocolo Chromium Genome HT Library & Gel Bead Kit V2. En éste, los fragmentos de ADN fueron asignados en una perla (GEM), en la cual fueron introducidos los códigos de barra (o barcodes) y los adaptadores, los fragmentos fueron amplificados en cada perla, posteriormente se realizó una fase de limpieza y se llevó a cabo la fase de reparación de los extremos seguido de la ligación de los barcodes para el secuenciamiento. La construcción

de la librería se completó con la selección de fragmentos y enriquecimiento por PCR. La librería con tamaños de inserto entre 500-700 bp fué cuantificada en un instrumento Qubit 2.0 fluorometer (Thermo Fisher Scientific, Waltham, MA, USA) y PCR cuantitativa. La distribución del tamaño de los fragmentos se analizó en un instrumento Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, EE. UU.). La libreria se secuenciaó en la plataforma Illumina HiSeq (Illumina, San Diego, CA, EE. UU.) generando lecturas pareadas de 150 pb.

Para el secuenciamiento con la tecnología PacBio, se construyó una librería empleando el SMRTbell Template Prep Kit, para el cual se usaron 10 ug de ADN de alta calidad. El ADN fue fragmentado aleatoriamente con un buffer de fragmentación mediante el uso de dispositivos Covaris, posteriormente se realizó la purificación de la muestra empleando perlas magnéticas AMPurePB. Se seleccionaron fragmentos de ADN superiores a 3 kb que pasaron por una fase de reparación de daño combinado con reparación de los extremos. Los extremos romos se ligaron con adaptadores tipo horquilla (o hairpin). El protocolo finalizó con una fase de purificación del ADN molde, anillamiento de cebadores y secuenciamiento en la plataforma PacBio Sequel.

En el presente estudio, la extracción del ADN de alto peso molecular, la preparación de las librerías Illumina, 10X y PacBio, y el secuenciamiento de las librerías fue realizado por la empresa Novogene (https://en.novogene.com).

3.5.2. Fases de pre-procesamiento, procesamiento y evaluación de calidad

Fase de pre-procesamiento

- Illumina y 10x Genomics:
  Empleando fastQC v.0.11.2 [1] se realizó la evaluación inicial de la calidad de las lecturas, identificando adaptadores y lecturas de baja calidad. El análsis de contenido de secuencias sobrerrepresentadas,(realizada con un k-mer de longitud seis) evidenció la necesidad de remover los extremos de las lecturas. Para ello se empleó el programa Trimmomatic v.0.36. [14], el cual hace uso de dos estrategias (simple y palindrome) para la identificación y remoción de adaptadores (Bolger et al, 2014). Los parámetros empleados para la búsqueda de adaptadores consideró un mismatch de 4 pb, una coincidencia de 24 pb en el modo palíndrome, una coincidencia de 9 pb entre adaptador - lectura y una longitud mínima de 120 pb de la secuencia evaluada.

- PacBio:
  Con los 25.7 Gb de secuenciamiento producidos con la plataforma de PacBio,se gene-

raron a partir de los archivos BAM, seis archivos en formato fasta, para ser empleados en la etapa de ensamblaje.

Fase de procesamiento

- Illumina y 10x Genomics:
  La etapa de ensamblaje se desarrolló en dos etapas. En la primera etapa se realizó el ensamblaje con el programa SOAPdenovo2 (127-mer) v.2.04 [72] de las secuencias de la librerias producidas por la tencologia Illumina, evaluandose seis tamaños de k-mers: 51, 61, 71, 81, 91 y 101.

  Teniendo como criterio el tamaño total del ensamblaje y el estadístico N50, se seleccionaron dos tamaños de k-mers: 71 y 81, los cuales fueron empleados en la segunda etapa de ensamblaje *de novo* donde se incorporaron las lecturas de la tecnología de 10x, sin los barcodes iniciales de las lecturas.

- PacBio:
  Se emplearon 2.156.844 lecturas en el ensamblaje *de novo* con el software Canu v.1.6. [67]. Posteriormente se mapearon las lecturas limpias de Illumina y 10x al ensamblaje obtenido a través del programa Bowtie2 v.2.3.3 [74] y se estimaron las estadísticas de calidad, cubrimiento y tamaño de fragmento a través del módulo QualStats del software NGSEP (Next Generation Sequencing Eclipse Plugin) v3.1.1 [32] [90]

Fase de evaluación de los ensamblajes

Para los diferentes ensamblajes se estimó el total de bases ensambladas (tamaño del ensamblaje), el número y el tamaño de scaffolds y contigs, y el estadístico N50. Adicionalmente, se evaluó el contenido génico de genes ortólogos de copia única, empleando la herramienta BUSCO v 3.0.2 [109]. La base de datos usada fue embryophyta(odb9), que contiene 1.440 genes de copia única de 30 especies de plantas. Para las predicciones iniciales de esta herramienta realizada con Augustus v 3.2.3 [97], se empleó como especie modelo *Arabidopsis thaliana*.

Adicionalmente, con el objetivo de evaluar el porcentaje de mapeo en cada uno de los ensamblajes, se empleó un conjunto de datos de GBS (genotyping-by-sequencing) de 10 accesiones (Tabla **3-4**) de frijol Lima, entre silvestres (W) y domesticadas (Dom) pertenecientes al acervo mesoamericano y andino.

Teniendo en cuenta las diferentes evaluaciones de los genomas ensamblados, se seleccionó el ensamblaje con los mejores resultados (ensamblaje-Pacbio), este se empleó como genoma de referencia al cual fueron alineados mediante bowtie2 [74] las librerías 1 y 2. A partir del

archivo de los alineamientos,mediante el software NGSEP 3.12 [90] se estimó la tasa de error, calculada a partir del número de diferencias con el genoma de referencia sobre el total de alineamientos únicos.

Tabla 3-4.: Accesiones silvestres (W) y domesticadas (Dom) de frijol Lima, pertenecientes al acervo mesoamericano y andino, con datos de GBS que se usaron para evaluar el porcentaje de mapeo en los ensamblajes.

Accesión	**Acervo**	**D/w**	**Accesión**	**Acervo**	**D/w**
G25108	Andino	Dom	G25228	Mesoamericano	W
G25540	Andino	Dom	G25410	Mesoamericano	Dom
G26459	Andino	W	G25787	Mesoamericano	Dom
G26468	Andino	W	G27345	Mesoamericano	W
G26617	Andino	Dom	G27455	Mesoamericano	Dom

En la figura **3-10** se detalla el proceso que se llevó a cabo para el ensamblaje del genoma de frijol Lima. El proceso inició con la fase (**A**) experimental de obtención de las muestras y su secuenciamiento (pasos 1:fase de campo, germinación de las semillas y obtención de trifolios. 2: Extracción de ADN. 3:Control de calidad de la muestra. 4:Construcción de librerías. 5: Secuenciamiento), seguido de la fase (**B)** de pre-procesamiento (pasos 6: Control de calidad. 7:limpieza de lecturas. 8:Contol de calidad 9:Lecturas de alta calidad para el ensamblaje) y la fase (**C)** de procesamiento (pasos 10-11: ensamblaje de lecturas, 12:Obtención de diferentes ensamblajes), y finalizó con la fase (**D)** de contol de calidad.

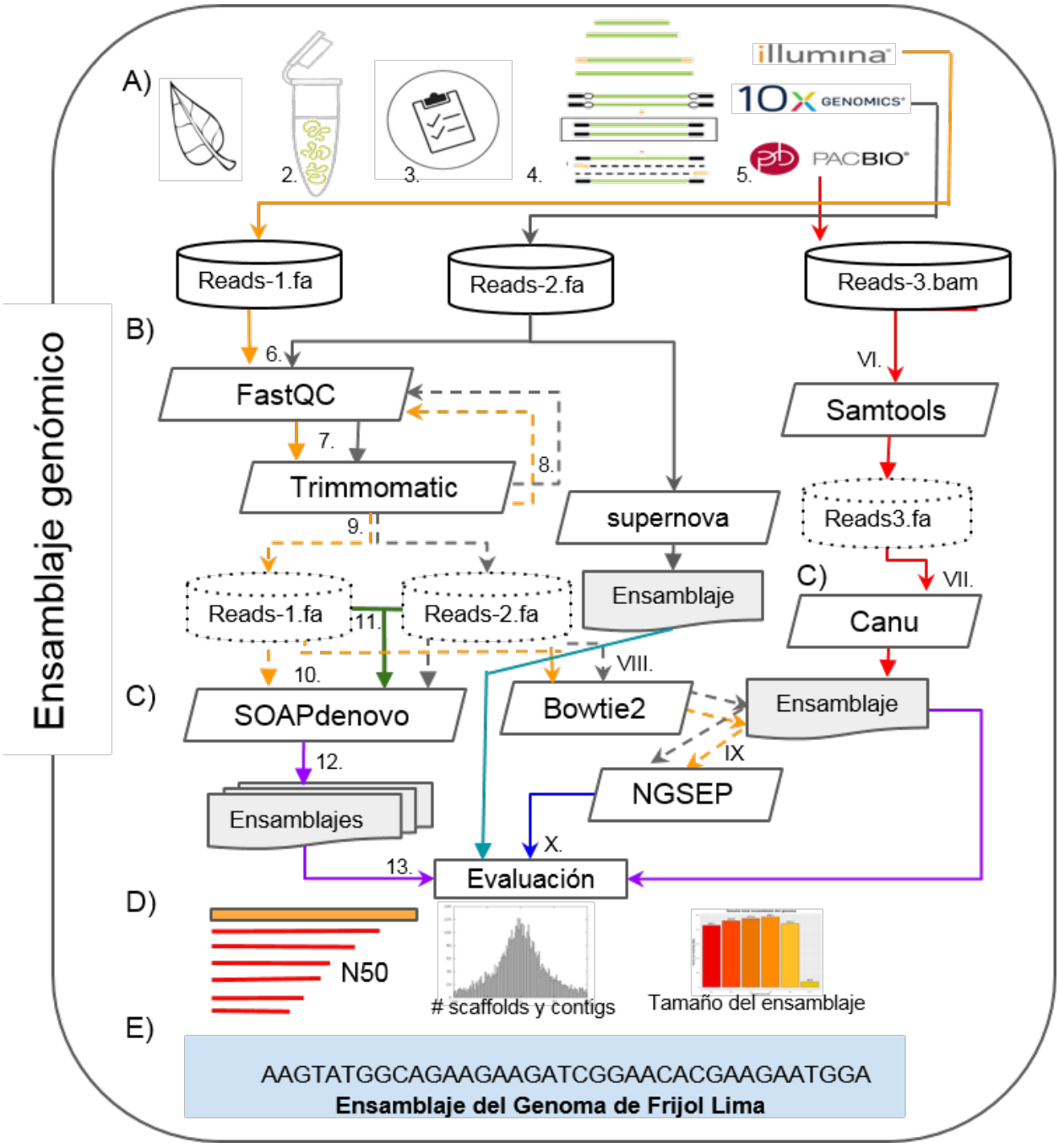


Figura 3-10.: Herramientas utilizadas en las diferentes fases del ensamblaje genómico **A)** Fase de producción de datos(1-5). **B)**Fase de pre-procesamiento(6-9). **C)** Fase de procesamiento-ensamblaje (10). **D)**Fase de evaluación de los ensamblajes(13). **E)** Genoma ensamblado.

4. Ensamblaje de novo de transcriptoma de frijol Lima

4.1. Resumen

El frijol Lima (*Phaseolus lunatus* L.) es una leguminosa tropical y subtropical, caracterizada por su diversidad y potencial de rendimiento pero que se cultiva a una escala mínima [58]. Su distribución abarca diferentes zonas ecológicas que comprende desde los cero a los 2.500 msnm, sugiriendo importantes adaptaciones ecológicas y transformaciones fenotípicas para responder a la diversidad de condiciones ambientales características de estas zonas. Bajo el anterior enfoque se evidencia la importancia de la caracterización estructural y funcional de los genes de frijol Lima.

En este estudio se reporta el secuenciamiento y ensamblaje *de novo* de alta calidad de los transcriptomas de tres tejidos, hoja, vaina y flor, del genotipo mesoamericano de frijol Lima domesticado G27455. Se secuenciaron 33.7 Gb de datos que fueron empleados en una estrategia tanto de ensamblaje guiado por referencia como *de novo*. En esta estrategia se evaluaron dos tamaños de k-mer y se generaron nueve ensamblajes, tres para cada tejido. Posteriormente, los ensamblajes de transcriptoma fueron evaluados y empleados en la anotación del genoma de frijol Lima con la identificación de 48127 genes. Con respecto a la anotación funcional se hallaron 2864 términos de ontologia de genes, de los cuales 1444 corresponden a la categoría de procesos biológicos, 343 a la categoría de componentes celulares y 1048 a la categoría de función molecular.

Conclusión: Se realizó la anotación estructural y funcional del genoma de frijol Lima, integrando evidencia experimental propia de la especie, mediante los ensamblajes de transcriptoma de tres importantes tejidos (hoja, flor y vaina).

Palabras claves: *Phaseolus lunatus*, transcriptómica, ensamblaje

4.2. Introducción

La familia de las leguminosas se destaca por su capacidad de fijar nitrógeno atmosférico en el suelo, a través de la relación simbiótica con bacterias del género *Rhizobium* y exudados de las raíces [86]. Sus frutos son fuente de un alto contenido de proteínas importantes en la dieta de humanos y como forrajes en la alimentación de animales [27]. Debido a su importancia agrícola se han desarrollado diversas investigaciones con el objetivo de caracterizar el genoma del mayor número de especies de leguminosas e identificar la expresión de genes implicados en características agronómicas, en diversos tejidos y etapas de desarrollo [43].

En el género *Phaseolus*, las investigaciones se han centrado en el frijol común para el cual se han reportado diversos ensamblajes de transcriptomas de diferentes tejidos como hojas, vainas, flores y raíces [62, 105]. Adicionalmente estos tejidos han sido evaluados en diferentes etapas de desarrollo (vegetativas y reproductivas) y bajo factores bióticos y abióticos [110, 7] para identificar genes diferencialmente expresados en las condiciones evaluadas. En cuanto a las investigaciones realizadas en frijol Lima, en el año 2015 se reportó el primer secuenciamiento y ensamblaje *de novo* del transcriptoma foliar del frijol Lima obtenido a partir del tratamiento con el hongo *Trichoderma viride*. En este estudio se identificaron genes involucrados en el tilacoide, el proceso de fotosíntesis y la generación de metabolitos, relacionando estos procesos con la evolución adaptativa de frijol Lima y frijol común [69] . El trabajo de Li et al. (2015) es pionero en el estudio de la genómica del frijol Lima aportando información importante sobre su transcriptoma, pero desde un enfoque de estrés biótico específico; por consiguiente, se hace necesario realizar nuevas investigaciones para el frijol Lima con el objetivo de caracterizar el transcriptoma de esta especie.

Las diversas investigaciones sobre transcriptomas de diferentes especies han sido posible gracias a los avances en las plataformas de secuenciamiento, especialmente a través de RNA-seq. Para el uso de esta tecnología a nivel experimental, se requiere la extracción del ARN mensajero a partir del ARN total, el cual es empleado como molde para la acción de la enzima transcriptasa reversa o retrotranscriptasa, para sintetizar ADN complementario de doble cadena que es empleado en el proceso de secuenciamiento [107]. Con esta tecnología se ha generado para varias especies un repositorio de datos conocidos como atlas de la expresión génica [86]. Adicional a la fase de obtención de datos de secuenciamiento, el análisis bioinformático requiere del ensamblaje de las lecturas y para ello existen dos enfoques: (1) ensamblaje *de novo* y (2) ensamblaje guiado por un genoma de referencia. El segundo enfoque es ampliamente usado cuando la especie cuenta con un genoma de referencia de alta calidad, lo cual permite el alineamiento de lecturas a este genoma y su empleo en la predicción de transcritos alternativos, también conocidos como isoformas [83].

Similar al proceso de análisis de datos de secuenciamiento de ADN, el ensamblaje *de novo* de

lecturas de ARN requiere de algoritmos que permitan reconstruir las secuencias completas a partir de la información de redundancia entre las diferentes lecturas. En la figura **4-1** se describe la metodología del software Trinity [48], el cual parte de una construcción de una tabla hash de las lecturas de acuerdo a un tamaño de k-mer y evalúa la abundancia de éste. El k-mer con mayor abundancia es empleado como semilla de extensión para generar contigs lineales. Posteriormente se realiza la construcción de una serie de grupos conformados por los contigs que comparten una superposición de K -1 entre ellos. A partir de estos grupos se construye un grafo de Bruijn donde cada nodo es el tamaño de palabra de k - 1, y k el eje del grafo. Finalmente se realiza una interacción entre los nodos del grafo de Bruijn para construir secuencias más largas y encontrar los caminos que son soportados por pares de lecturas en el grafo generando los diferentes contigs de ensamblaje [48].

Para el caso de ensamblajes de genomas nuevos, los resultados del ensamblaje *de novo* de lecturas de RNA-seq se pueden utilizar como evidencia experimental para la identificación de genes. Para ello, diversos procesos de anotación usan este recurso que, combinado con predicciones *ab initio*, generan la caracterización estructural de genes en los genomas. Esta estrategia está implementada en el software Maker, el cual integra en sus cinco etapas herramientas de predicción *de novo* de genes, identificación de regiones repetitivas, integración de datos experimentales, consenso de evidencias y corrección final. La integración de diferentes fuentes de información permite obtener una anotación estructural de genes de alta calidad [18].

En la presente investigación se generaron datos de secuenciamiento de RNA-seq de frijol Lima para tres tejidos diferentes (hoja, vaina, flor), con los cuales se realizaron dos ensamblajes *de novo* y uno guiado por referencia para cada tejido. Estos ensamblajes fueron empleados en la anotación del genoma donde se identificaron 48127 genes, los cuales se clasificaron en 2864 términos de ontología, 1444 de ellos en la categoría de procesos biológicos, 343 en la categoría de componentes celulares y 1048 en la categoría de función molecular.

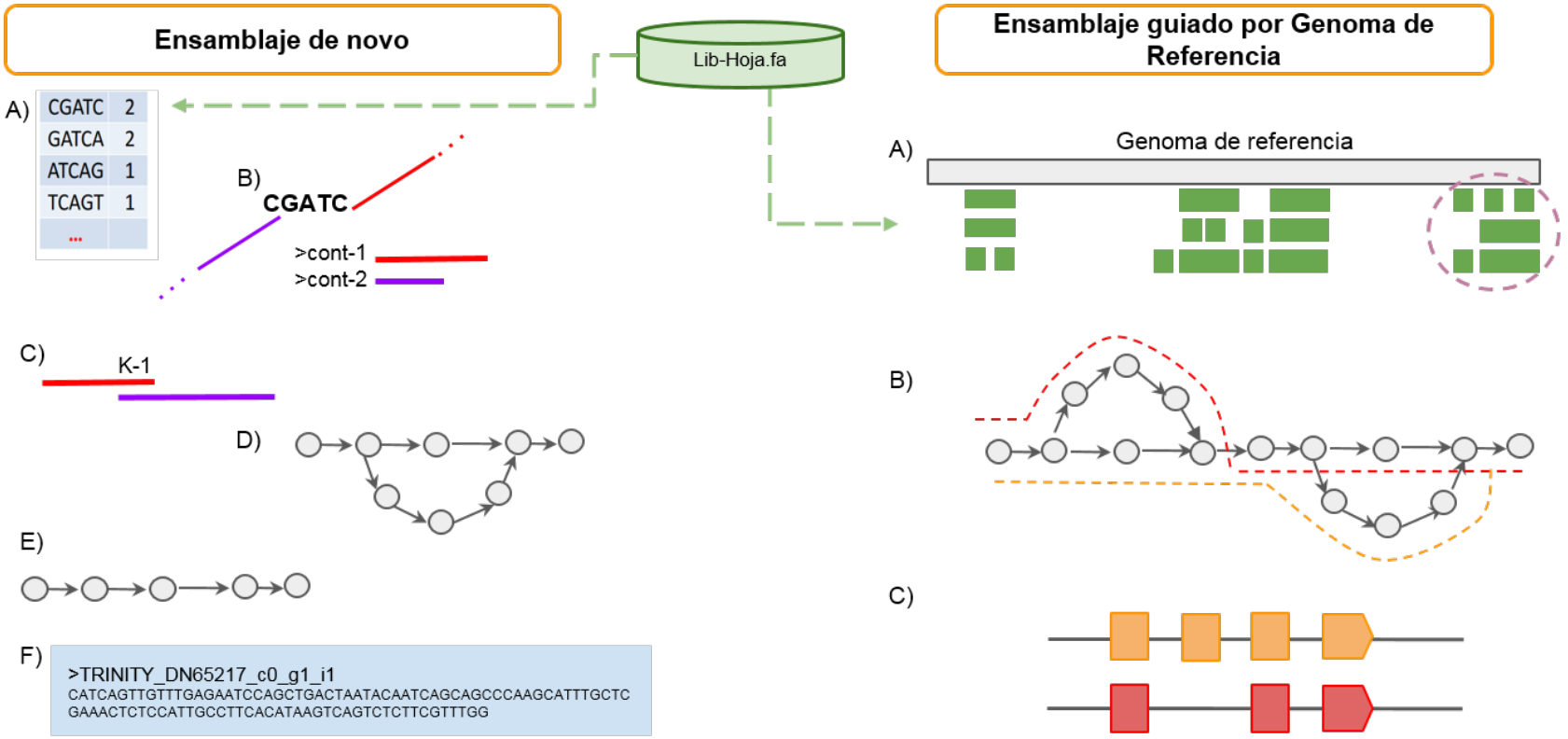


Figura 4-1.: Ensamblaje *de novo* y por referencia con líbrerias de RNA-seq implementado en Trinity

En el ensamblaje *de novo* se distinguen cinco etapas: **A)**Contrucción de tabla hash de acuerdo acuerdo a un tamaño de secuencia (kmer) **B)**Uso del kmer con mayor abundancia como semilla de extensión **C)** Construcción de contigs lineales **D)**Construcción del de Bruijn **E)**Interacción entre los nodos del grafo de Bruijn **F)** Reporte de los diferentes transcritos ensamblados. En el ensamblaje guiado por un genoma de referencia, se identifican tres etapas: **A)** Alineamiento de las lecturas de RNA-seq al genoma **B)** Construcción de un grafo en cada región donde se obtienen evidencia de múltiples lecturas **C)**Proceso de iteración sobre el grafo para reconstruir los transcritos.

4.3. Resultados y Discusión

4.3.1. Conjunto de datos de secuenciamiento

Siguiendo el protocolo de RNA-seq se geneneraron a través de la plataforma Illumina datos de secuenciamiento de ARN para tres tejidos de frijol Lima. En total se generaron 111 millones de lecturas pareadas, que contenían 16.6 Gb. Para la librería de hoja se obtubieron 23 millones de lecturas (6.9 Gb de datos crudos), para la librería de flor 19 millones (5.9 Gb de datos crudos) y para la librería de vaina 69 millones de lecturas (20.7 Gb de datos crudos). En la tabla **4-1** se describen los rendimientos obtenidos en cada líbreria.

Tabla 4-1.: Producción de datos de secuenciamiento de RNA-seq obtenidos por la tecnologia Illumina a partir de tres librerías (hoja, flor y vaina) en el frijol Lima.

Libreria	Plataforma	Longitud promedio de lectura (pb)	Número de lecturas (M)	Bases totales (Gb)	Datos crudos (Gb)
Hoja			23	3.4	6.9
Flor	Illumina	150	19	2.9	5.9
Vaina			69	10.3	20.7
		Total	111	16.6	33.5

[1] pb: pares de bases, M: millones, Gb: gigabase.

4.3.2. Evaluación de calidad de las lecturas

Al realizar la evaluación de las lecturas se encontró que para las tres librerías la calidad por base fue bastante alta con un valor >30.(Anexo B, fig:**B-1**, fig:**B-3** y fig:**B-5**). Sin embargo, a través del contenido de k-mer (Anexo B, fig:**B-2**, fig:**B-4** y fig:**B-6**). se identificaron secuencias sobrerepresentadas en los extremos de las lecturas, indicando la presencia de adaptadores. Se removieron el 1.03 %, 1.37 % 0.77 % de los datos iniciales para la libreria de hoja, flor y vaina, respectivamente.

4.3.3. Estrategía de ensamblaje de novo del transcriptoma de Frijol Lima

La estrategía de ensamblaje empleada (figura **4-2**) consideró dos enfoques: ensamblaje *de novo* y ensamblaje guiado por referencia (empleando como genoma de referencia el ensamblaje de alta calidad descrito en el capitulo anterior). Para los ensamblajes *de novo* se usaron dos tamaños de k-mer (25 y 31), como tamaños extremos posibles de usar en el software de ensamblaje. De acuerdo a la anterior estrategía se generaron dos ensamblajes *de novo* y uno

guiado por referencia para cada uno de los tejidos evaluados para un total de nueve ensamblajes. En la tabla **4-2** se encuentra las caracterización inicial de cada uno de los ensamblajes en cuanto al número de transcritos, bases ensambladas, tamaño ensamblado, entre otras.

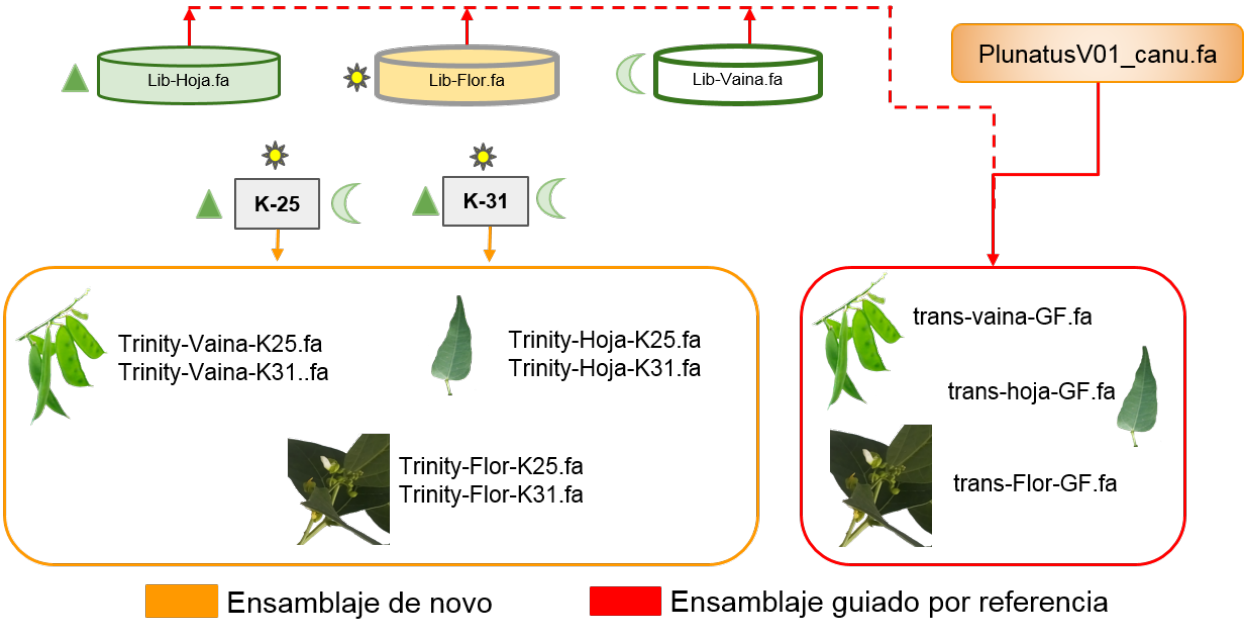


Figura 4-2.: Estrategía de ensamblaje de transcriptomas *de novo* usando dos tamaños de kmer (25 y 31) y guiado por la referencia.

Tabla 4-2.: Métricas primarias de los ensamblajes de transcriptoma obtenidos a partir de los tejidos de hoja, flor y vaina en frijol Lima.

Tejido	Ensamblaje	Total transcritos	Total bases ensambladas (Mb)	% GC	contig N50
Hoja	K-25	163197	178	41.32	1747
	K-31	153634	172	41.33	1798
	GF	124344	138	40.84	1765
Flor	K-25	167450	165	41.55	1542
	K-31	158419	159	41.55	1585
	GF	114231	115	40.95	1548
Vaina	K-25	246774	259	40.65	1678
	K-31	234378	254	40.62	1725
	GF	165165	169	39.81	1658

[1] GF: ensamblaje guiado por referencia

[2] Mb: megabase.

De los nueve ensamblajes generados, se observa que el mayor número de transcritos reportados se obtiene con un tamaño de k-mer de 25 pb, donde el tejido de la vaina contó con 246.774, seguido del tejido de la flor (167.450) y la hoja (163.197). En contraposición a esta tendencia, el ensamblaje guiado por la referencia reportó el menor número de transcritos en todos los tejidos. El número de transcritos reportados está relacionado con el número de bases ensambladas con los valores más altos para el ensamblaje de K-25. Con respecto al N50 en los tres tejidos, el mayor fue para K-31. Al considerar esta estadística para cada tejido se observan valores muy cercanos en los tres ensamblajes, con una discrepancia de aproximadamente 50 pb. En cuanto al porcentaje de GC, éste fluctuó entre 39 y 41 %, con valores muy similares en todos los ensamblajes.

4.3.4. Evaluación de los ensamblajes de transcriptoma

Los nueve ensamblajes obtenidos se evaluaron a través de tres mediciones. La primera medición consistió en el porcentaje de lecturas que alinearon a cada una de las librerías de RNA-seq en el ensamblaje respectivo. La segunda medición fue la cuantificación del número de ortólogos de copia única, y la tercera medición fue el número de transcritos con la longitud más larga del conjunto de proteínas evaluadas con el proteoma de frijol común.

En cuanto a la primera medición (fig **4-3**) se encontró que para todos los ensamblajes la tasa de alineamiento fue superior al 90 % indicando que la mayoría de las lecturas de las librerías fueron empleadas en el proceso de ensamblaje. De acuerdo a Haas et al. (2014) se puede considerar un buen ensamblaje cuando el porcentaje de alineamiento de lecturas es superior al 80 %. Es importante destacar que en los tres tejidos, la menor tasa de alineamiento se obtuvo para el ensamblaje guiado por referencia y la mayor tasa para el ensamblaje con un tamaño de k-mer de 31.

Con el objetivo de evaluar la integridad de los ensamblajes (fig **4-4**), éstos fueron comparados con un conjunto de 1.440 ortólogos de copia única altamente conservados en plantas [114]. Se halló que el ensamblaje del tejido de hoja empleando un k-mer de 25 fue el que obtuvo el mayor porcentaje de ortólogos con 86.3 % (11.7 % para genes de copia única y 74.6 % para genes duplicados), seguido por el tejido de la vaina con 84.3 %, y el tejido de la flor con 78 %. Del conjunto de ensamblajes, el porcentaje de genes ortólogos más bajo estuvo para el ensamblaje k-31 del tejido de la flor con 77 %. Al comparar los genes reportados para cada ensamblaje, se encuentra que 1.189 genes ortólogos fueron compartidos entre todos los ensamblajes. Sin embargo, este valor incluye genes duplicados. Al considerar solamente el número de ortólogos de copia única se observa que los ensamblajes K-25 en los tres tejidos contienen el mayor número de genes, 169 (11.7 %) en hoja, 218 (15.1 %) en flor y 161 (11.2 %) en vaina. Para todos los ensamblajes, los ortólogos perdidos oscilaron entre el 9 %

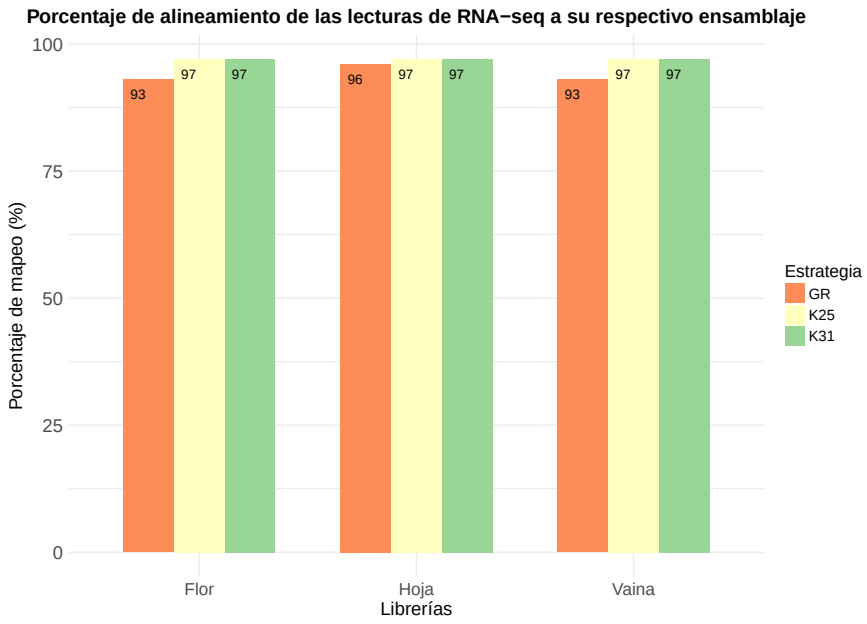


Figura 4-3.: Tasa de alineamiento de lecturas de RNA-seq a los diferentes ensamblajes de transcriptoma

y el 15 % con un total de 103 genes ortólogos compartidos. Con respecto al porcentaje de fragmentación, el más alto fue para el tejido de la flor en los tres ensamblajes, y el más bajo para el tejido de la hoja. De acuerdo a este críterio de evaluación se evidenció una buena cobertura de genes en los ensamblajes, donde el k-mer de tamaño 25 presenta los mejores desempeños.

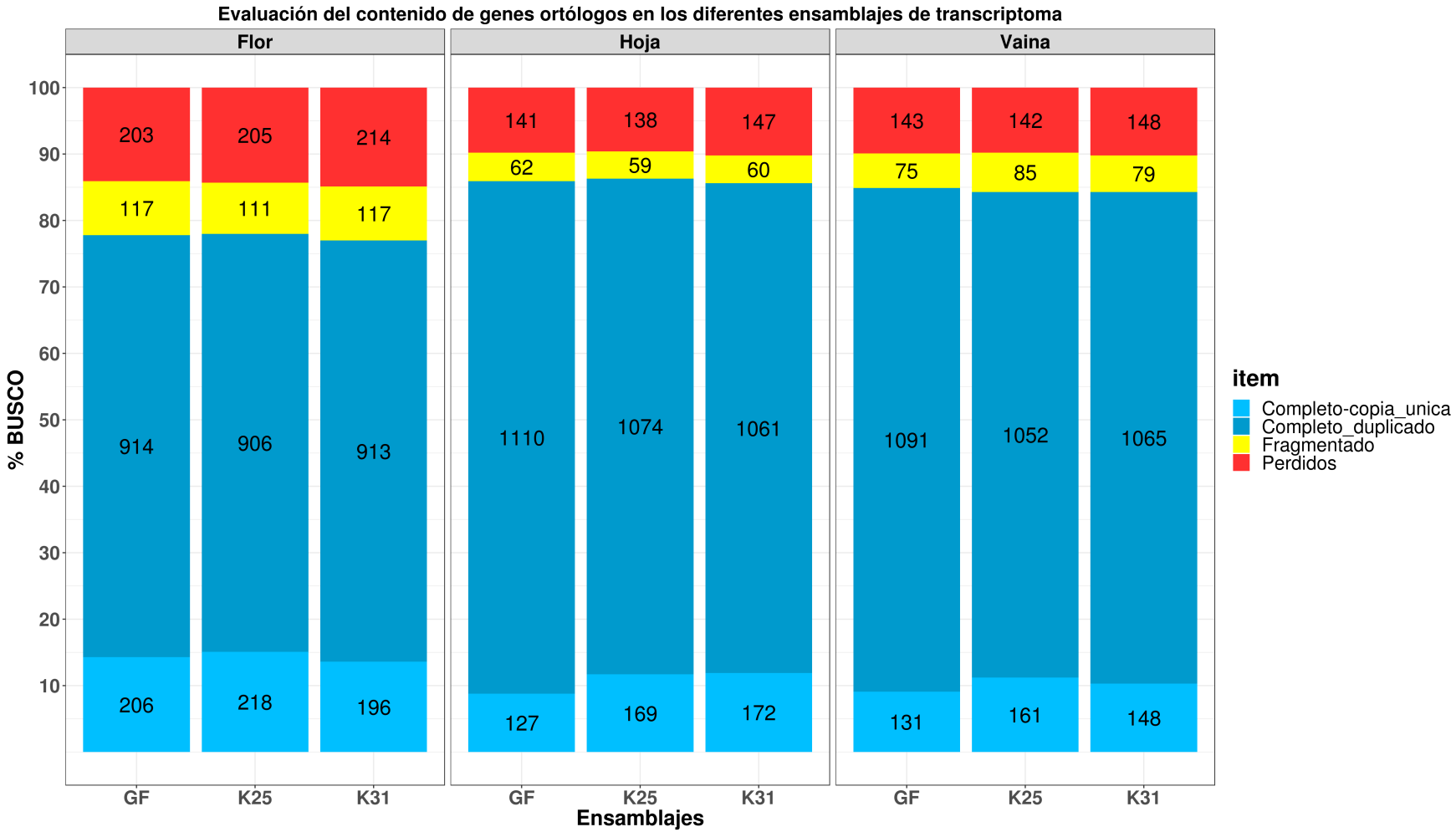


Figura 4-4.: Número de genes ortólogos encontrados en los diferentes ensamblajes de transcriptoma de frijol Lima para los tejidos de flor, hoja y vaina, ensamblados con la estrategia *de novo* (con k-mers de 25 y 31) y guiado por referencia (GF).

Los ensamblajes también se evaluaron comparandolos contra el catálogo de 36.995 proteinas extraido del genoma de frijol común versión 2 (fig **4-5**). Para alrededor del 60 % de los transcritos obtenidos en los diferentes ensamblajes se encontró un ortólogo potencial en frijol común con un cubrimiento de su secuencia superior al 70 %. Nuevamente, los porcentajes de transcritos encontrados más altos se encontraron en los ensamblajes guiados por referencia (74 % para el ensamblaje de las lecturas de vaina). En contraste, los menores porcentajes se encuentran para los ensamblajes con k-25.

De acuerdo a las métricas empleadas, se puede evidenciar que los ensamblajes *de novo* del transcriptoma con los dos tamaños de k-mer empleados (25 y 31) no presentan una gran diferencia, debido a la pequeña variación en el número de bases (6 pb) del kmer. No obstante, en la evaluación del número de genes ortólogos de copia única tuvo un mejor desempeño el k-mer 25. En cuanto al ensamblaje guiado por referencia, en las tres evaluaciones sus resultados estuvieron muy cercanos al k-mer 25, e incluso se encontró en la reconstrucción de longitudes de transcritos un número mayor de coincidencias. Por lo anterior, se considera una buena alternativa la combinación de estrategias de ensamblaje (*de novo* y guiado por referencia) para contribuir en la identificación del contenido génico de frijol Lima.

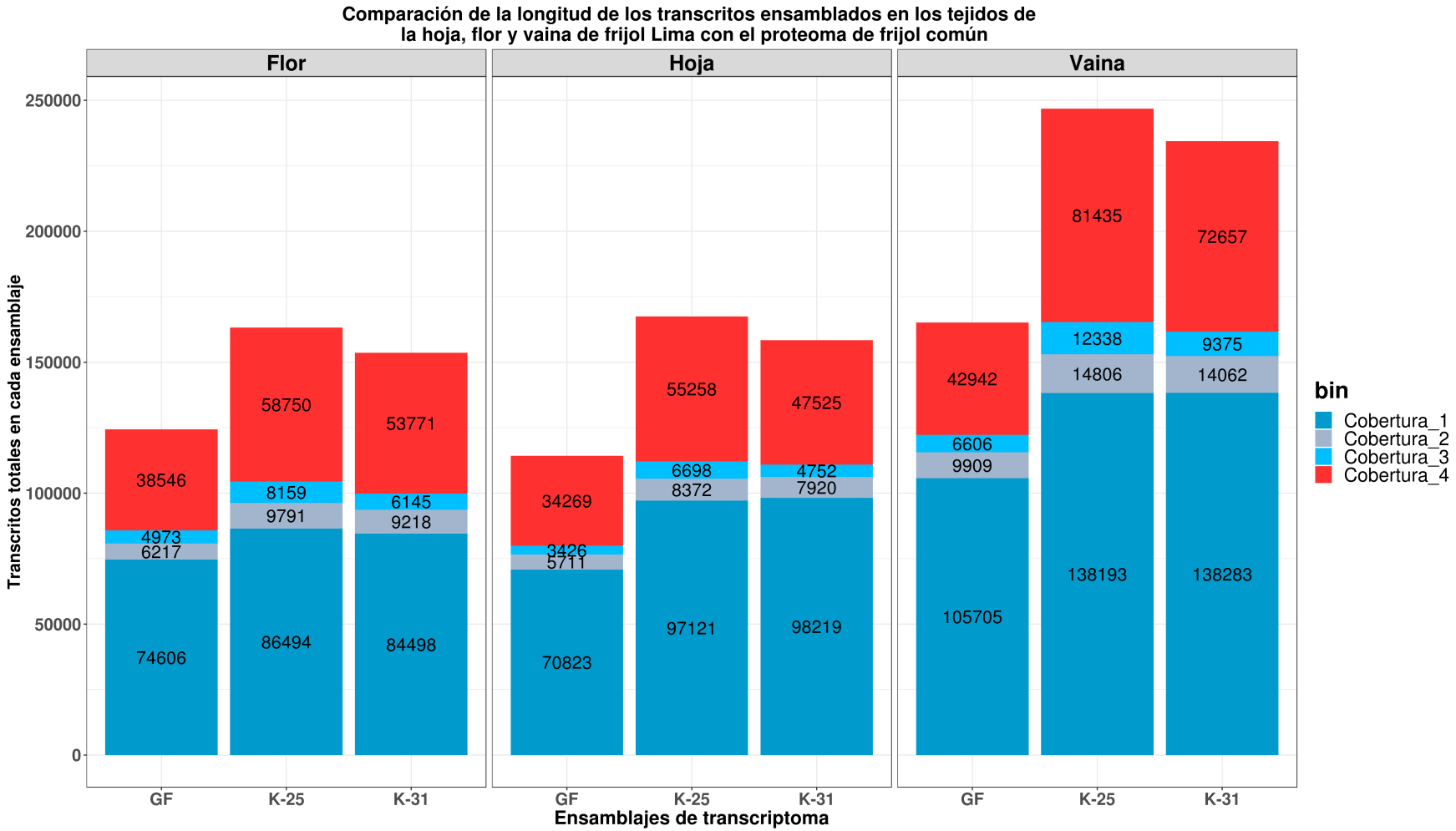


Figura 4-5.: Comparación de la longitud de los transcritos ensamblados en los tejidos de la hoja, flor y vaina de frijol Lima con el proteoma de frijol común.

En la gráfica La cobertura 1 representa los transcritos con una cobertura >90 % y <=100 %, la cobertura 2 indica una cobertura >80 % y <=90 %, la cobertura 3 comprende el rango entre >70 % y <=80 % y la cobertura 4 representa un cubrimiento <70 % (color rojo).

4.3.5. Anotación del genoma de frijol Lima

Aproximación para identificar regiones repetitivas en el genoma de frijol Lima

Empleando 6.180 elementos repetitivos, se enmascaró el 17.50 % del genoma de frijol Lima, es decir 94 millones de bp. Se encontró que 2.49 % son repeticiones simples de tipo microsatélite. De los 496 contigs del genoma, se observó que el mayor número de bases enmascaradas se encuentran en los contigs n°13, 1848, 1846,29 y 1856, y la menor en los contigs n°669, 1988, 2062, 2074 y 792. No obstante, al comparar el porcentaje de elementos repetitivos obtenidos para frijol lima, con los reportados para frijol común con el 41 % del genoma, se evidencia una gran diferencia, posiblemente se debe a que las secuencias repetitivas pueden ser específicas de especie o género [76] y para frijol común se usó un conjunto de repeticiones identificadas por el grupo de Scott A. Jackson [51], propias de frijol común. En nuestro caso se empleó el conjunto de repeticiones de la base de datos dcotrep [103], con elementos repetitivos comunes a las plantas con flores, dejándose fuera del análisis posibles elementos repetitivos más informativos propios del género *Phaseolus*, pero no disponibles en su totalidad en bases de datos públicas. Adicionalmente se caracterizó el tamaño de las regiones repetitivas en el genoma de frijol Lima, las cuales aproximadamente abarcan 30 % del genoma ensamblado para frijol Lima. En la gráfica 4-5 se observa que el 16 % del total de las regiones repetitivas presentan un tamaño entre 400 a 599 pb, seguido del 14 % con un tamaño entre 200-399 y un 12 % con un tamaño de 600 a 759. Este primer avance en la caracterización de las regiones repetitivas para frijol Lima, evidencia la importancia de estas regiones en la organización del genoma de esta especie y sugiere la necesidad de determinar en futuras investigaciones su incidencia a nivel evolutivo y funcional.

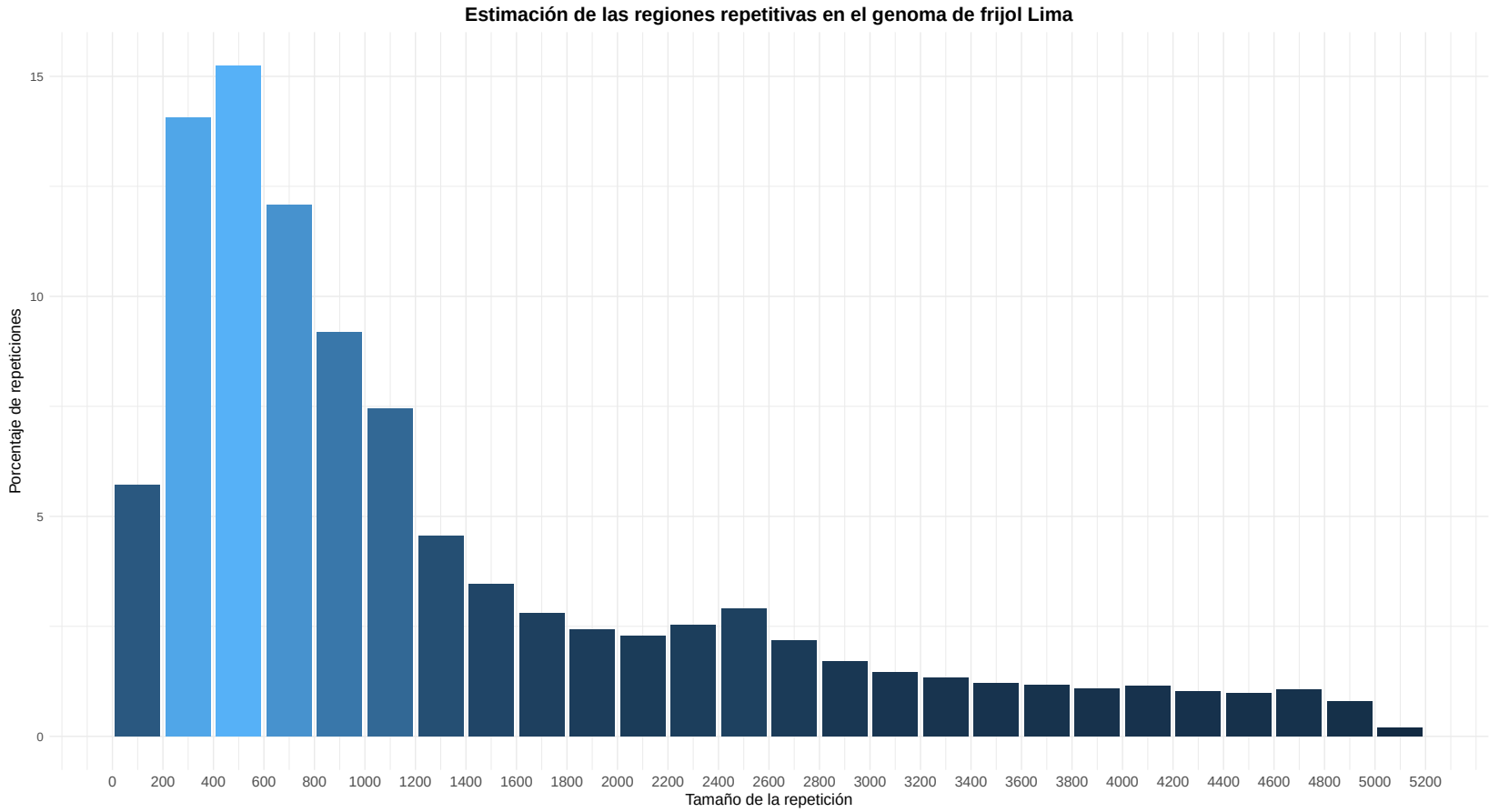


Figura 4-6.: Estimación del tamaño de las regiones repetitivas en el genoma de frijol Lima

Anotación Estructural

Se identificaron 48.127 genes para frijol lima, los cuales en su mayoría tienen un único transcrito (47.644), 393 tienen dos transcritos y 19 tienen 4 transcritos (fig **4-7**). Al comparar esta tendencia con frijol común en las dos versiones del genoma se encuentra que fríjol Lima tiene un comportamiento similar, puesto que la mayoría de genes reportan un único transcrito, 21564 para frijol común v2.1, 22311 para frijol común v1.0 y 32806 para arroz (*Oryza sativa*). En cuanto al número máximo de transcritos reportados por gen se encuentra que frijol común v 2 tiene 20 transcritos para dos genes, mientras que en la versión 1 del genoma de esta especie el máximo número de transcritos es de 13. Igual que en arroz, en frijol Lima el mayor número de transcritos por gen son 6, con un solo gen con esta característica.

Al continuar con el análisis de los genes de frijol Lima se observa que el tamaño de los genes, como se observa en la fig **4-8** , se ubica en un rango entre 200 y 400 pb, esta misma tendencia se observa en *O. sativa*, sin embargo para frijol común en las dos versiones del genoma el tamaño se ubica entre 1200 y 1400 pb. Se identifica que frijol Lima presenta una similaridad en los tamaños grandes de los genes de frijol común a partir del rango de 1200 a 1400 pb, mientras que en los tamaños pequeños frijol Lima se comporta como *O. sativa*.

En cuanto al número de exones presente por transcrito (fig **4-9**), el frijol Lima presenta un 24 % del total de los transcritos con un único exón, mientras que *O. sativa* presenta el 14 % con 3 exones, al igual que frijol común v1 con 15 % y en la versión 2 con 10 % de los transcritos con 5 exones. Se observa que el frijol Lima presenta la mayor cantidad de transcritos con uno y cuatro exones. Con respecto a la longitud de los transcritos (fig.**4-10**), el frijol Lima presenta la mayor cantidad de transcritos en el rango entre 200 y 400 pb con 7638, esta tendencia es similar en *O. sativa* con 4980 en esta categoría, mientras que en frijol común (v1.0 y v2.1) el mayor punto de datos se encuentra entre 1200-1400 con 2624 y 3576 transcritos respectivamente.

Al realizar la estimación de la longitud de las proteínas (fig. **4-11**) se observa que el mayor número de éstas se agrupan en longitudes inferiores a 200 aminoácidos, en frijol Lima y en arroz, mientras que en frijol común v1 y v2 se encuentra entre 200 a 400 aminoácidos. En los bines de agrupación posteriores a 600 aminoácidos, las tres especies tienen una tendencia similar con valores inferiores a las 3500 proteínas con dicha longitud.

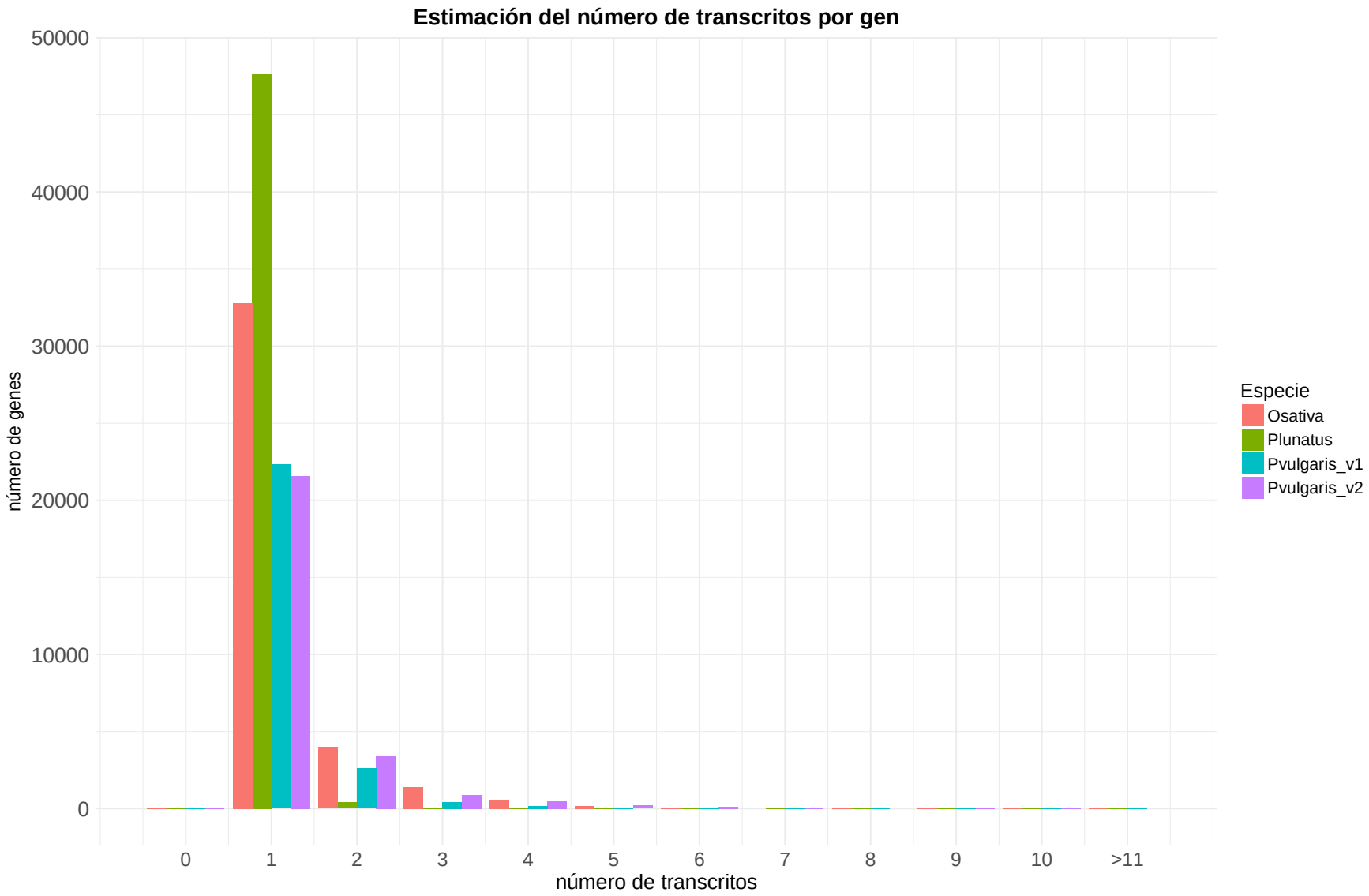


Figura 4-7.: Estimación del número de transcritos por gen identificados en la anotación estructural del frijol Lima

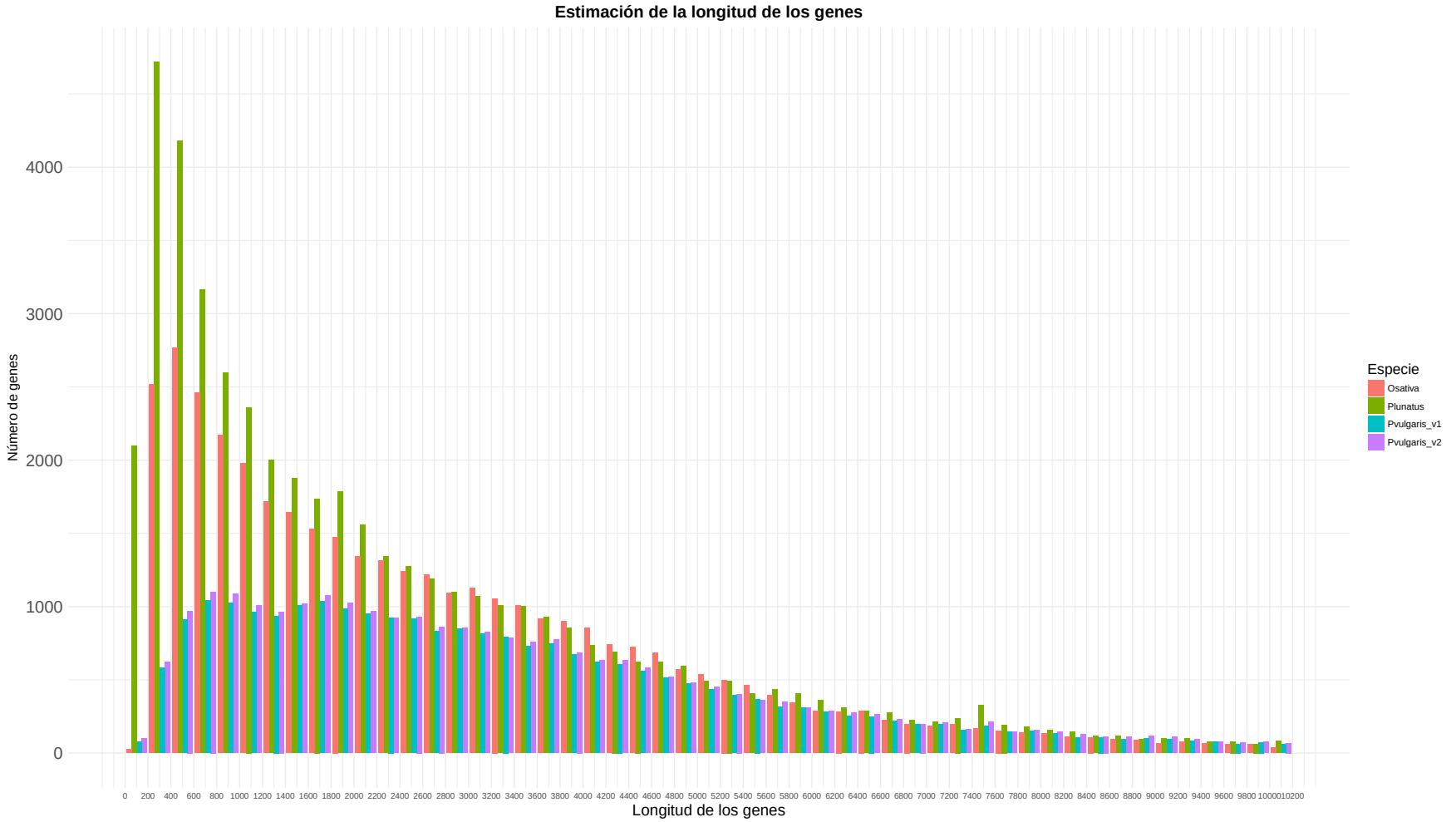


Figura 4-8.: Estimación tamaño de los genes identificados en la anotación estructural del frijol Lima.

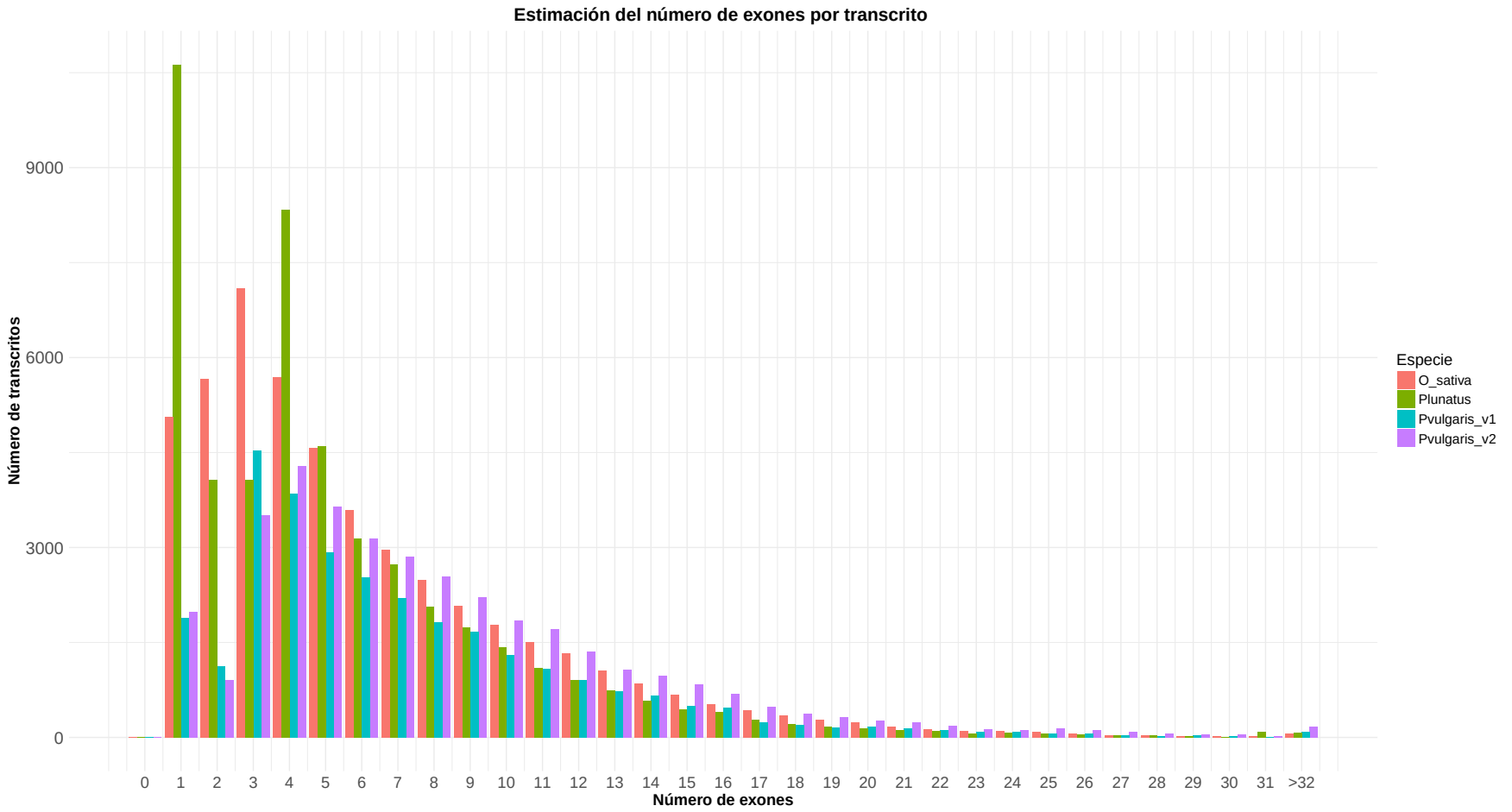


Figura 4-9.: Estimación del número de exones por transcritos identificados en la anotación estructural del frijol Lima.

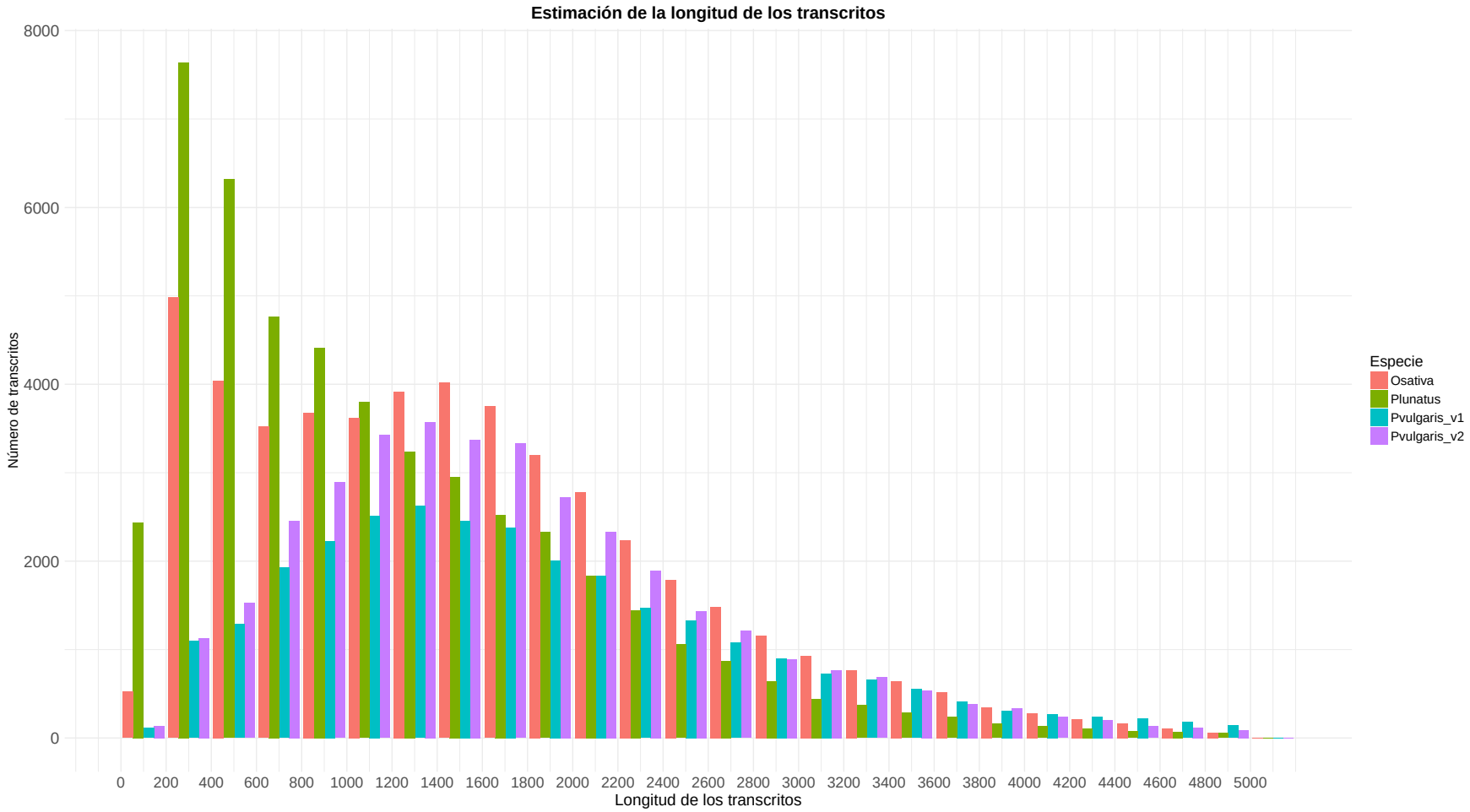


Figura 4-10.: Estimación de la longitud de los transcritos identificados en la anotación estructural del frijol Lima

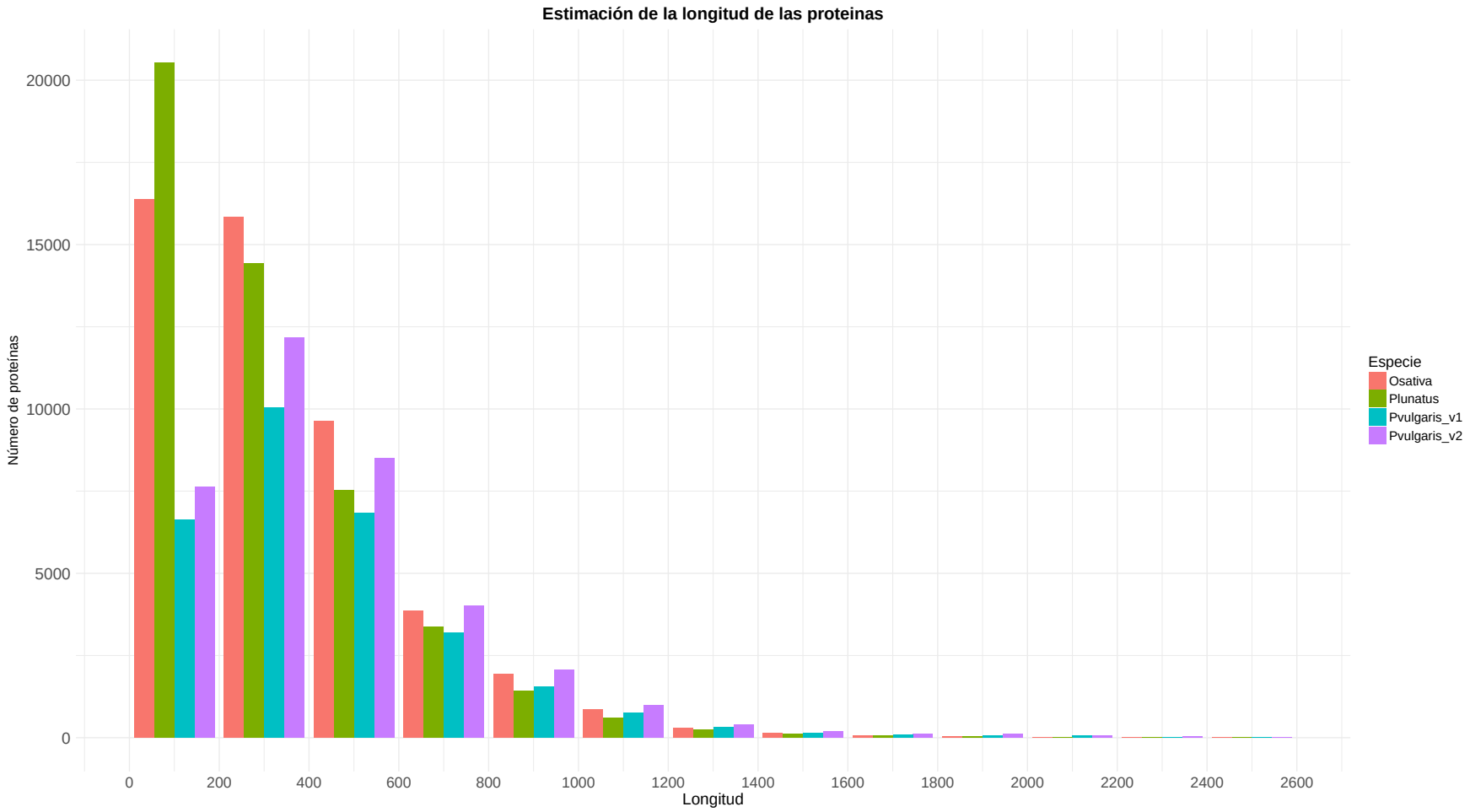


Figura 4-11.: Estimación de la longitud de las proteínas identificadas en la anotación estructural del frijol Lima

Anotación Funcional

De las 54.311 proteínas identificadas en la fase de anotación estructural del genoma de frijol Lima, se obtuvieron 18.448 coincidencias entre el proteoma de frijol Lima y la base de datos de UniProt (http://www.uniprot.org/), de éstas se mantuvieron 5.148 al considerar un porcentaje de similitud del =>70 %, como lo sugiere Schmutz (2014) [98] en la anotación para frijol común. A partir de las 5.148 coincidencias se obtuvieron 3.828 identificadores únicos de proteínas, los cuales tienen asociados 26.775 términos de ontología (gene ontology). Al reducir la redundancia de estos términos se obtuvieron 2.864 registros únicos. De éstos, 1.444 corresponden a la categoría de procesos biológicos, 343 a la categoría de componentes celulares y 1.048 a la categoría de función molecular.

Cada una de las anteriores categorías se comparó con el porcentaje de anotaciones de ontología reportadas para *Arabidopsis thaliana*, con relación a la categoría de procesos biológicos como se observa en la fig **4-12**. Se encuentra que los mayores porcentajes de genes se ubican en la anotación de procesos celulares (72 %), seguido de la anotación de procesos metabólicos (60 %) y regulación biológica (22 %). Un ejemplo de genes en estas anotaciones son respectivamente: APC6 - anaphase-promoting complex subunit 6, NUA poro nuclear y NEDD1 - transducin/wd40 domain-containing protein.

En cuanto a la categoría de componentes celulares (fig **4-13**), se encuentra que el 80 % de los genes hacen parte de la anotación de parte celular y complejo molecular, seguido con el 75 % de proceso celular y el 62 % del proceso metabólico. Algunos genes asociados a parte celular son: membrana celular B1386G10.7, regiones de la vacuola Os01g0221600 y citoesqueleto LOC100785401. En la categoría de función molecular (fig **4-14**) se identifican genes asociados a la anotación unión (64 %) y actividad catalítica (50 %) que actúan sobre el ARN como SIG2 (rna polymerase sigma), AT1G26370 (rna helicase family protein) y NRPA2 (nuclear rna polymerase a2).

Al comparar las anotaciones de ontología de *Arabidopsis thaliana* con respecto a frijol Lima se observa que los mayores porcentajes de genes reportados en *Arabidopsis thaliana* coinciden con los máximos porcentajes de frijol Lima. Sin embargo, la tendencia es que frijol Lima reporta mayores porcentajes, esto se debe a la diferencia en genes anotados para *Arabidopsis thaliana* con 27.655 [51], y posiblemente a la diferencia en el tamaño de los genomas.

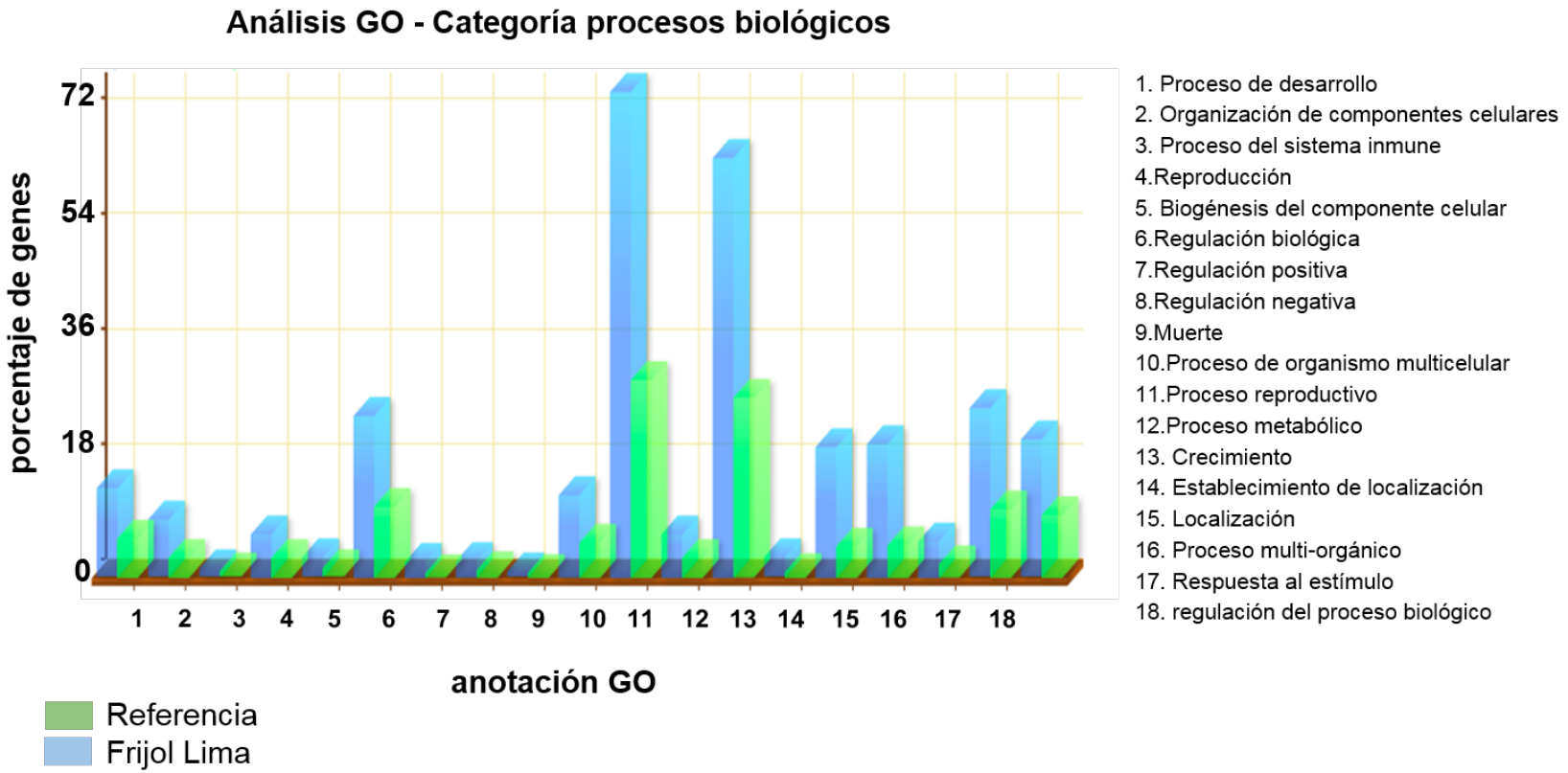


Figura 4-12.: Comparación de la categoría GO procesos biológicos de frijol Lima con respecto a *Arabidopsis thaliana*. Al lado derecho de la gráfica se encuentra cada una de las categorías GO con su respectivo identificador.

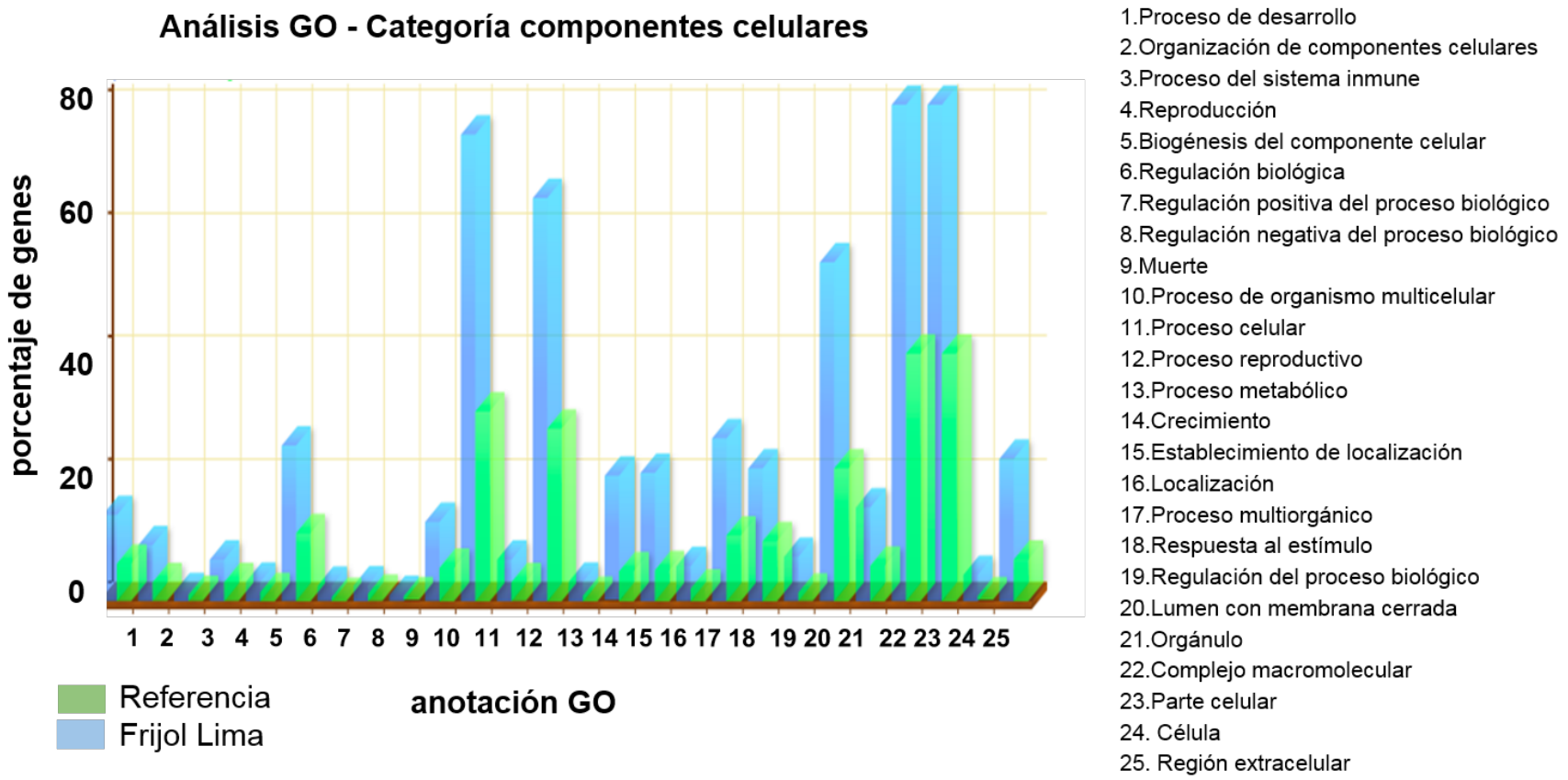


Figura 4-13.: Comparación de la categoría GO componentes celulares de frijol Lima con respecto a *Arabidopsis thaliana*. Al lado derecho de la gráfica se encuentra cada una de las categorías GO con su respectivo identificador.

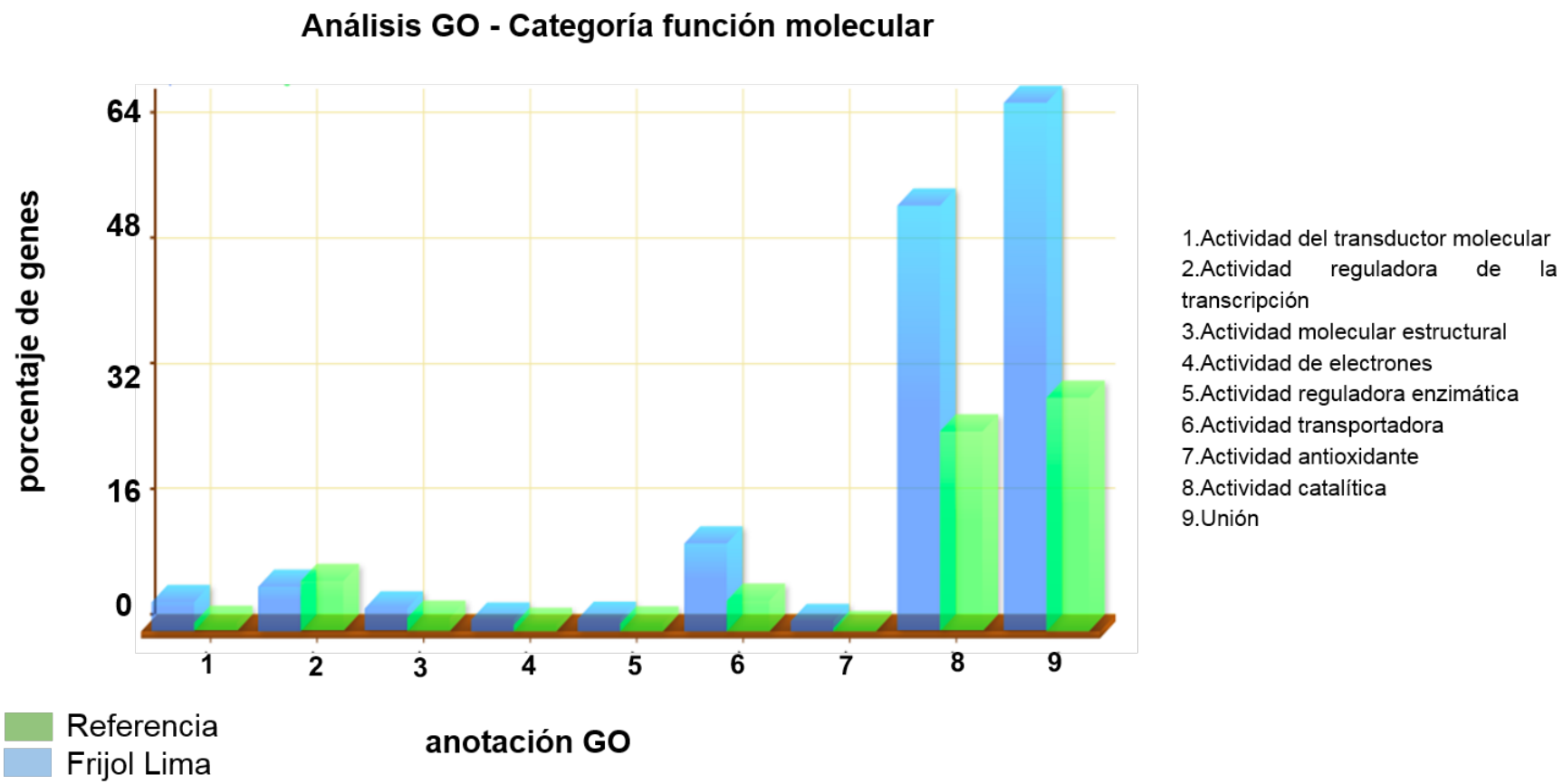


Figura 4-14.: Comparación de la categoría GO función molecular de frijol Lima con respecto a *Arabidopsis thaliana*.
Al lado derecho de la gráfica se encuentra cada una de las categorías GO con su respectivo identificador.

4.4. Conclusiones

Se generaron nueve ensamblajes de transcriptoma, dos *de novo* y uno guiado por la referencia, para tres tejidos (hoja, flor y vaina) de frijol Lima. Se observó que el k-mer de tamaño 25 presentó las mejores mediciones de evaluación en cuanto al porcentaje de lecturas mapeadas a cada ensamblaje, número de genes ortólogos de copia única y el número de transcritos con la longitud más larga.

Mediante la integración de evidencia experimental (ensamblajes de transcriptomas) y predicción *de novo* se identificaron 48127 genes en el genoma de frijol Lima, los cuales presentan en su mayoría (47644) un único transcrito. Adicionalmente se identificaron 54311 proteínas, con una longitud promedio de 200 aminoácidos. El análisis de ontología de genes relaciona 2864 términos GO, de éstos 1444 corresponden a la categoría de procesos biológicos, 343 a la categoría de componentes celulares y 1048 a la categoría de función molecular.

4.5. Materiales y métodos

4.5.1. Obtención de ARN, construcción de librerías y secuenciamiento

El ARN total fue obtenido de tres tejidos (hoja joven, flor y vaina) de la accesión G27455 (Sucre-Colombia) del acervo mesoamericano de frijol Lima. El protocolo de extracción fue específico para cada tejido. Para el tejido de la vaina se empleó el protocolo modificado de Wang et al, (2009) [107], (modificaciones sin publicar). Para el tejido de la flor se utilizó el protocolo de la doctora Adriana Bohórquez (sin publicar) del CIAT y para el tejido de la hoja un protocolo modificado a base de trizol del doctor Maruyama (sin publicar).

Las muestras de ARN fueron secadas en SpeedVac empleando RNAlater Invitrogen, y enviadas a secuenciamiento a la empresa Novogene en Sacramento (California-USA). Las librerías fueron enriquecida con perlas oligo (dT), posteriormente el ARN ribosomal fue removido con el kit Ribo-Zero. El mRNA fue fragmentado y empleado para la construcción del cDNA, adicionalmente los extremos fueron reparados y ligados a los adaptadores. Finalmente el cDNA fue sometido a una selección por tamaño y enriquecido por PCR. El control de calidad fue realizado en tres momentos: en primer lugar se evaluó la concentración de la muestra a través de un instrumento Qubit 2, posteriormente mediante un instrumento Agilent 2100 se evaluó el tamaño del inserto y finalmente a través de Q-PCR se cuantificó la concentración de las muestras. El secuenciamiento se realizó en la plataforma Hiseq de Illumina con lecturas pareadas de 150 pb.

4.5.2. Fases de pre-procesamiento, ensamblaje y validación de los ensamblajes

Se generaron 33.7 Gb de datos de secuenciamiento entre las tres librerías, éstas fueron evaluadas con el software fastQC v.0.11.2 [1], para identificar adaptadores en los extremos de las lecturas y regiones de baja calidad en las primeras 25 pb del extremo 5'. La estrategia de limpieza de datos se realizó con el software Trimmomatic v.0.36 [14] considerando un mismatch de 4pb, una coincidencia de 24pb en el modo palíndrome, una coincidencia de 9 pb entre adaptador - lectura, un headcrop de 15 pb y una longitud mínima de 110 pb de la secuencia evaluada.

El ensamblaje se realizó a través de una estrategia con dos enfoques: ensamblaje *de novo* y guiado por referencia. Se empleó el software Trinity v2.4.0 [48] en los dos casos. Para el ensamblaje *de novo* se usaron dos tamaños de k-mer: 25 y 31, y para el ensamblaje guiado por referencia se empleó el genoma de frijol Lima caracterizado en el capítulo tres de este trabajo. Para este último ensamblaje, se realizó previamente el mapeo de las lecturas de cada librería al genoma con Hisat2 v1.0.1 [63], lo que produjo un archivo bam que se empleó como entrada principal en el ensamblaje. De acuerdo con la estrategia de ensamblaje anteriormente descrita, se obtuvieron nueve ensamblajes, los cuales se evaluaron en tres aspectos. El primer aspecto es el porcentaje de alineamiento de las lecturas al transcriptoma ensamblado. El segundo, es la longitud de los transcritos, con el objetivo de hallar el número de transcritos que coinciden con al menos $> 70\,\%$ con proteinas de frijol común versión 2.1, el cual cuenta con 36.995 proteínas [51]. Para ésto se empleó la herramienta Blastx v 2.2.28 [23] con un e-value de 1e-20. Y el tercer aspecto fue la cuantificación de la integridad del conjunto de datos genómicos a través del número de genes ortólogos de copia única presente en cada ensamblaje. Para ésto se empleó un conjunto de 1.440 genes presentes en las plantas terrestres (Embryophyta) mediante la herramienta BUSCO v 3.0.2 [109].

4.5.3. Identificación de regiones repetitivas

La identificación, clasificación y enmascaramiento de elementos repetitivos, incluidas las secuencias de baja complejidad y repeticiones intercaladas, se realizó con el software RepeatMasker v 4.0.5 [103], el cual contiene una librería de regiones repetitivas cuyas entradas se alinean al genoma. Se emplearon 6.180 elementos repetitivos de la librería de plantas con flores (dcotrep version 23.03), con los parámetros -e ncbi -s -a -x. Adicionalmente se empleó el software NGSEP (Next Generation Sequencing Experience Platform) v3.1.2 [90] para identificar regiones repetitivas.

4.5.4. Anotación del genoma

La anotación estructural se llevó a cabo mediante el software Maker, el cual realiza la integración de datos experimentales, homología y predicciones *ab initio* [18, 17]. Para ésto, se incorporaron los tres ensamblajes de los transcriptomas que se ensamblaron con el k-mer de tamaño 25, el proteoma de frijol común version 2 y el genoma de frijol Lima. La predicción *de novo* se realizó con las herramientas snap, gmhmme3 y augustus, empleando la especie *Arabidopsis* como modelo. Para el análisis de la anotación estructural se desarrollaron scripts en java usando clases del software NGSEP v3.12. [32].

Empleando el archivo de anotación estructural generado con Maker, se realizó la anotación funcional en cuatro etapas:

- La primera etapa consistió en análisis de similitud de secuencias a nivel de proteínas, mediante Blastp v 2.2.29 [23], con un evalue de 1e-20, contra la base de datos Uniprot (release 2018-02) [22].
- En la segunda etapa, la anotación primaria fue filtrada teniendo como criterio un porcentaje de identidad mayor al 70 % con las coincidencias encontradas.
- Un segundo filtro se empleó en la tercera etapa, con el objetivo de verificar que las proteínas que cumplieron el anterior filtro pertenecieran a la categoría taxonómica de plantas de acuerdo a la base de datos Taxonomy (release 3/12/18) del NCBI [23].
- En la cuarta etapa se realizó una curación manual descartando especies de hongos, catalogadas conjuntamente en la categoría de plantas reportada por la base de datos Taxonomy.

El archivo generado después de los filtros se empleó para obtener los términos de ontología génica (GO), mediante script desarrollados en bash y perl. Los términos reportados en cada una de las categorías GO, fueron comparados con la anotación GO de *Arabidopsis thaliana*. En la figura **4-15** se detalla el proceso que se llevó a cabo para el ensamblaje de los transcriptomas que inició con la fase experimental de obtención de las muestras y su secuenciamiento (fase A), seguido del procesamiento *in silico* (fase B). En la fase (C) se llevó a cabo el ensamblaje. Posteriormente los ensamblajes fueron evaluados y empleados en la fase de anotación representada con el número 11 en la gráfica.

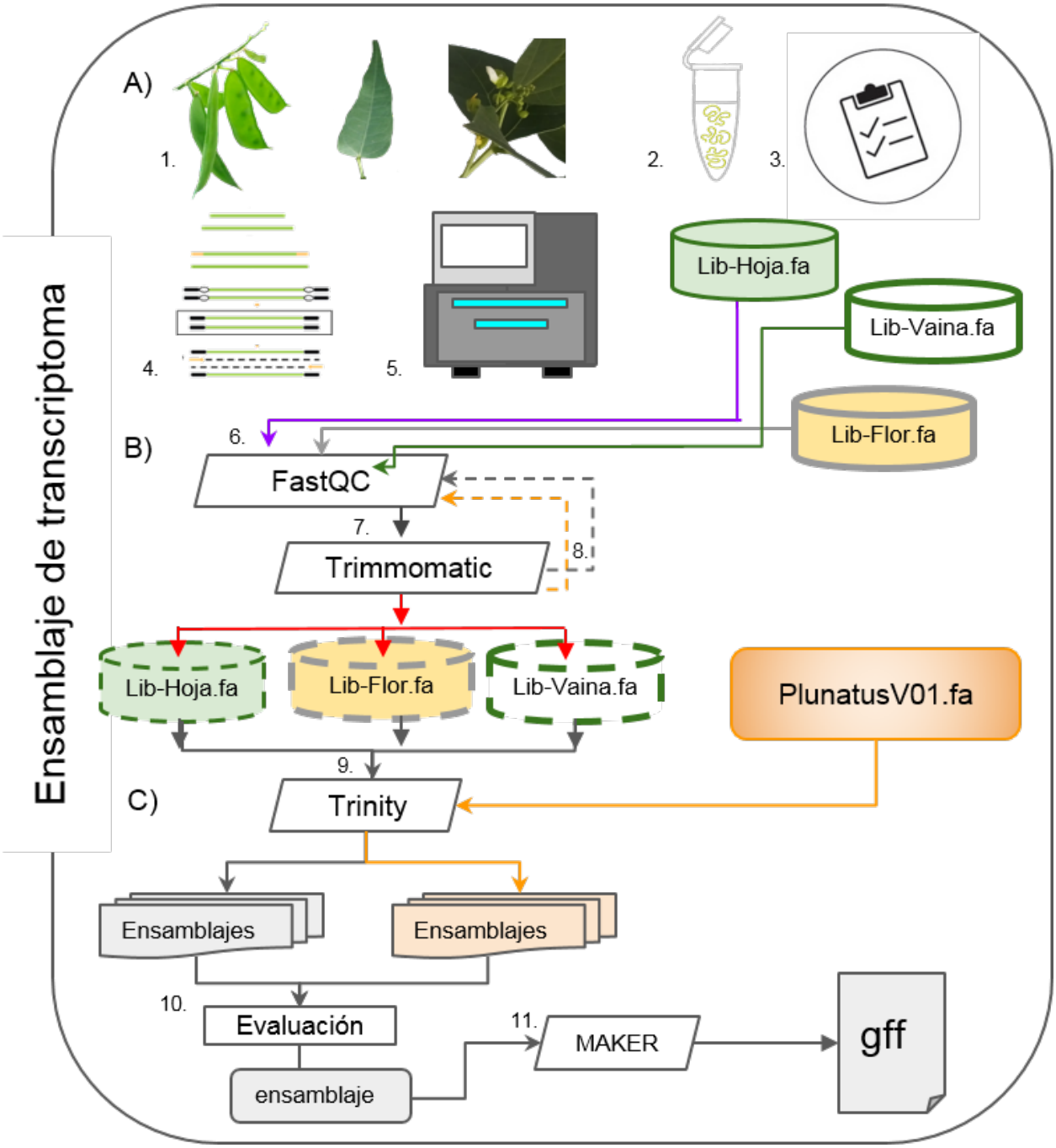


Figura 4-15.: Herramientas empleadas para el ensamblaje de transcriptoma y anotación del genoma.
(A) Fase experimental(1-5). **B)**Fase de pre-procesamiento(6-9) **C)** Fase de procesamiento (10) Fase de evaluación (11) Fase de anotación

5. Identificación de regiones microsinténicas asociadas a genes de domesticación en Frijol Lima

5.1. Resumen

El rápido desarrollo de las tecnologías de secuenciación ha conducido a la generación, cada vez más creciente, de ensamblajes de genomas de múltiples especies vegetales, cuyo análisis en un contexto comparativo ha permitido la identificación de regiones conservadas entre las especies [91] y la evaluación del orden, conservación y estructura de los genes [20]. Uno de los beneficios más grandes que ha traído la genómica comparativa es poder traducir la información funcional del genoma de una especie bien estudiada (o modelo) en otra especie con menos información, en especial la detección de genes ortólogos que puedan controlar rasgos fenotípicos de interés. *Phaseolus lunatus* L. (frijol Lima) y *Phaseolus vulgaris* L. (frijol común) son las especies más importantes del género *Phaseolus*, sin embargo estudios comparativos entre estas dos especies no habían sido posible debido principalmente a que hasta la fecha no se contaba con un ensamblaje del genoma de frijol Lima. El genoma generado para frijol Lima en la presente investigación provee las bases para la detección de genes que puedan estar controlando rasgos de interés agronómico a través de un enfoque comparativo.

En objetivo del presente trabajo fue detectar bloques microsinténicos entre frijol Lima, frijol común y frijol mungo asociados a genes de la domesticación, en específico al rasgo dehiscencia de la vaina. Este rasgo en cultivos de leguminosas es indeseable ya que causa grandes pérdidas en el rendimiento, por lo tanto el conocimiento de su control genético puede favorecer futuros programas de mejoramiento en estas especies. *Arabidopsis thaliana* ha sido la planta modelo para el estudio de los genes implicados en la dehiscencia de la vaina, lo que ha permitido la caracterización de cinco genes: *SHATERPROOF* (*SHP1*), *INDEHISCENT* (*IND*), *ALCATRAZ* (*ALC*), *FRUITFULL* (*FUL*) y *NST1* (*NAC SECONDARY WALL THICKENING PROMOTING FACTOR 1*). Estos cinco genes se usaron como punto de partida para la caracterización de regiones microsinténicas respecto al grado de conservación, contenido génico, grado de ordenamiento y orientación de los genes, con el fin de detectar en frijol Lima genes candidatos para el rasgo dehiscencia de la vaina.

Conclusión: Los genes *ALC*, *FUL* y *NST1* fueron los más conservados entre frijol Lima y frijol común en cuanto a su estructura (número de exones y longitud total de exones e intrones); para estos tres genes frijol mungo tuvo mayores diferencias y siempre presentó menor longitud total de los exones. Los genes *SHP1* e *IND* presentaron diferencias entre las tres especies en todos los aspectos estructurales evaluados. Se identificaron entre 1 y 7 bloques colineales en cada una de las regiones sub-genómicas (longitud de 200 Kb) donde se ubicaron los cinco genes de interés. En general, en las regiones subgenómicas no se observaron rearreglos entre los bloques colineales detectados, con excepción de una inversión en las regiones asociadas a los genes *IND* y *NST1*. Lo anterior indica un alto grado de conservación en las regiones subgenómicas que contienen estos cinco genes.

Palabras claves: domesticación, ortología, dehiscencia, microsintenia.

5.2. Introducción

El rápido desarrollo de las diferentes tecnologías de secuenciación, el aumento en el tamaño de la lecturas, la disminución de las tasas de error y la reducción en los costos de producción por Gb, han incrementado el secuenciamiento de genomas de especies no modelo [36], lo que ha permitido ensamblajes de alta calidad con su respectiva anotación de un gran número de especies, y con ello la oportunidad de comparar genomas completos (macrosintenia) o a menor escala regiones genómicas de interés (microsintenia) [20, 113]. Los estudios de genómica comparativa han permitido evaluar la conservación, localización y orden de genes ortólogos [104], es decir, genes derivados de un ancestro común [91].

A nivel bioinformático, la identificación de genes ortólogos es un reto debido a las distintas fuerzas evolutivas que actúan sobre los genomas, evidenciadas a través de eventos genéticos (duplicación de genes o regiones *en tandem*, deleción o nacimiento de genes) y rearreglos cromosómicos (inversiones, translocaciones, fisiones y fusiones) [68], que dificultan la detección de estos genes. No obstante, incluso entre especies separadas por millones de años de evolución y divergencia, aún existen regiones conservadas en los cromosomas que reflejan la organización ancestral de las especies, las cuales se pueden identificar mediante un enfoque comparativo[28] y cuya estructura se asume que deriva de un proceso de especiación desde un ancestro común. Las regiones cromosómicas donde el orden de genes es conservado entre las especies se conocen como regiones sinténicas. No solo es de interés establecer si el orden de los genes es conservado o no entre las especies, sino también establecer si los genes son ortólogos (copias de genes derivadas de una copia génica ancestral presente en el último ancestro común entre las especies comparadas) o parálogos (copias de genes derivadas por duplicación dentro de una especie)[64]. La evidencia experimental indica que los genes ortólogos tienden a retener funciones equivalentes entre las especies, aunque con

excepciones [64], por lo que es necesario validar experimentalmente su función. En este sentido, la identificación de regiones microsinténicas entre las especies puede proveer un fuerte soporte para establecer cuáles copias son putativamente ortólogas y favorecer el proceso de construcción de regiones macrosintenicas. Por su importancia, se han desarrollado diferentes métodologías para la identificación de copias ortólogas, las cuales se describen a continuación.

Los métodos de inferencia de genes ortólogos se basan en similaridad de la secuencia, filogenia y conservación sinténica. El primer enfoque supone que, al provenir de un ancestro común, los genes ortólogos son más parecidos entre sí cuando se comparan sus secuencias (de ADN o de proteínas) que con otros genes presentes en el mismo genoma. Por lo tanto, el mejor acierto recíproco (best reciprocal hit) que se obtiene al hacer una búsqueda por similtud de un gen de una especie en genes de otra especie debería corresponder con el ortólogo del gen buscado [64]. Este enfoque es el más simple y es más apropiado para genomas cercanamente relacionados, aunque para especies distantemente relacionadas puede ser también útil [64]. Descubrir genes ortólogos en diferentes especies es una actividad importante para realizar análisis filogenéticos ya que la filogenia construída a partir del alineamiento de ortólogos (gene tree) se puede comparar con la taxonomía de las especies (species tree), ya sea para reconciliarlas [64] o para establecer hipótesis sobre el origen de sus diferencias. El enfoque de conservación sinténica considera la conservación de los genes en orden y orientación a lo largo de los cromosomas, suponiendo que la cantidad de eventos de rearreglo estructural es relativamente pequeño entre especies cercanas. Este enfoque ha sido asociado con el enfoque de similaridad e integrado a la evaluación de los genes colindantes al gen ortólogo en estudio [102]. Para abordar el problema de identificación de relaciones de ortología, diversas herramientas han sido desarrolladas empleando elementos de los tres métodos de identificación de genes ortólogos descritos anteriormente, y modificaciones a éstos con el objetivo de mejorar la precisión, la velocidad y la especificidad en la identificación de relaciones de ortología en el conjunto de datos evaluados [85].

Como se detalló anteriormete, la identificación de genes ortólogos permite realizar de manera eficiente la anotación funcional de los genes identificados en la construcción de un nuevo genoma. Esto facilita la predicción de genes que pueden estar involucrados en el control genético de un rasgo fenotípico en la especie de interés. En las plantas cultivadas, la identificación de genes candidatos para rasgos agronómicos se ha realizado en parte a través de la identificación de genes ortólogos con la especie vegetal modelo *Arabidopsis thaliana* o con otras especies modelo. Aunque la predicción de función por ortología no es completamente confiable, realizar esta predicción contribuye información para seleccionar genes para experimentos de validación funcional, ya que el hecho de compartir ancestria con otras especies soporta la posible funcionalidad del gen.

La capacidad de dispersión de semilla es uno de los rasgos más interesantes relacionados con

el proceso de domesticación en plantas. En contraste con las plantas silvestres, cuyo éxito evolutivo depende de la capacidad para dispersar sus semillas y así garantizar su supervivencia, para las plantas domesticadas la dispersión de semillas es un rasgo no deseado que se redujo o se perdió durante su adaptación a los nuevos agroecosistemas. Esta tendencia se reporta no solo en las especies del género *Phaseolus*, sino también durante la domesticación de cultivos de semillas en general [5, 53]. En los cultivos de leguminosas, la apertura de las vainas permite que las semillas escapen a la cosecha humana por lo que la reducción de la dehiscencia del fruto es un rasgo de interés agronómico susceptible de mejora genética en estas especies. La identificación de los genes que controlan este rasgo en los cultivos de leguminosas es por lo tanto de especial interés.

Arabidopsis thaliana ha sido la planta modelo para estudiar la cascada de genes implicados en la dehiscencia de la vaina, lo que ha permitido la caracterización de cinco genes: *SHATERPROOF* (*SHP1*), *INDEHISCENT* (*IND*), *ALCATRAZ* (*ALC*), *FRUITFULL* (*FUL*) y *NST1* (*NAC SECONDARY WALL THICKENING PROMOTING FACTOR 1*). En la parte superior de las cascada reguladora se encuentran dos genes relacionados a la familia de factores de transcripción MADS-box, *SHATERPROOF* (*SHP1*) y *SHATERPROOF2* (*SHP2*) [70], los cuales controlan la diferenciación de la zona de dehiscencia (DZ) del fruto en *Arabidopsis* [35]. Actuando aguas abajo de los genes *SHP1 / 2* están los genes *IND* y *ALC*, el primero incide en la diferenciación de células en la DZ [73] y el segundo es necesario para la formación de una capa celular especializada no lignificada dentro de la DZ [93].

Adicionalmente, se requiere el gen *FUL* para la expansión y diferenciación de las valvas del fruto. *FUL* es un regulador negativo de *SHP1 / 2*, que restringe la expresión de *SHP1 / 2* e *IND* a la DZ [42]. El gen *NST1* actúa aguas abajo de *SHP1 / 2*, se expresa en la capa lignificada de células en la DZ y su mutación produce frutos indehiscentes [81]. Finalmente, para que tenga lugar la dehiscencia de la silicua en *Arabidopsis*, es indispensable la acción de la enzima conocida como endo-poligalactunorasas (PG) que degrada la pectina para promover la separación celular. Los genes *ARABIDOPSIS DEHISCENCE ZONE POLYGALACTONURASE1* (*ADPG1*) y *ADGP2* codifican las PG que se expresan específicamente en la DZ [87].

En la presente investigación, se utilizaron los cinco genes (*SHP1, IND, ALC, FUL* y *NST1*) reportados para el rasgo de dehiscencia de la vaina como punto de partida para la caracterización de regiones micro-sinténicas entre frijol Lima, frijol común y frijol mungo, con el fin de detectar en frijol Lima genes candidatos para el rasgo dehiscencia de la vaina. Las relaciones de microsintenia se establecieron respecto al grado de conservación, contenido génico, grado de ordenamiento y orientación de los genes que comparten el ambiente cromosómico donde se ubican estos cinco genes de la domesticación.

5.3. Resultados y Discusión

5.3.1. Caracterización de los datos genómicos

Para realizar las comparaciones sinténicas se emplearon tres especies: frijol Lima, frijol común y frijol mungo. El criterio de selección de estas especies fue evolutivo, considerándose el clado *Phaseoleae* [19], donde el fríjol común y el fríjol Lima son especies cercanas que pertenecen al grupo B del clado *Phaseolus* propuesto por Delgado-Salinas et al. (2006), sin embargo se ubican en grupos diferentes. En el grupo vulgaris se encuentra *P. vulgaris* y sus especies relacionadas principalmente de Mesoamérica, mientras que en el grupo lunatus se incluyen solamente las especies de Sur América como *P. lunatus* y especies endémicas de diferentes islas como *P. mollis* Hook.f. de Galápagos [31]. En cuanto al frijol mungo, esta especie pertenece al género *Vigna* que es hermano del clado *Phaseolus* [19].

A nivel genómico, frijol Lima, frijol común y frijol mungo presentan el mismo número de cromosomas (2n=22), sin embargo el grado de calidad del ensamblaje genómico es diferente, asi como también, el tamaño estimado del genoma, el tamaño de ensamblaje y el número de genes identificados en cada uno. En la tabla **5-1** se presenta el resumen de las características principales de estos genomas donde se observa que el genoma de frijol Lima se ensambló en el presente estudio en 541 Mb, 496 contigs y un N50 de 5.5 Mb. En cuanto al genoma de frijol común versión 2.1, el tamaño ensamblado fue de 537 Mb en 11 cromosomas, 467 scaffolds y un N50 de 49.7 Mb [98, 51]. El genoma del frijol mungo fué ensamblado en 463 Mb, organizado en 11 cromosomas y 2488 scaffolds, y un N50 de 1.52 Mb[61].

Tabla 5-1.: Caracterización de los genomas de frijol Lima, frijol común y frijol mungo

Especie	Tamaño estimado del genoma (Mb)	Tamaño ensamblado (Mb)	Número de genes	Referencia
Frijol Lima	600	541	48127	Este trabajo
Frijol común	600	537	36995	[98]
Frijol mungo	579	463	22427	[61]

5.3.2. Caracterización de la región subgenómica asociada al gen SHATERPROOF (SHP1)

SHATERPROOF (*SHP1*) hace parte de la familia de los factores de transcripción tipo MADS-box [70]. En el genoma de frijol común, este gen (id Phvul.006G169600) se ubica en la cadena negativa del cromosoma 6, tiene un tamaño de 7.656 pb y ocho exones. En frijol mungo, este gen se encuentra con un porcentaje de identidad del 94 % (id Vradi10g09690)

con respecto a frijol común en la cadena negativa del cromosoma 10, presenta un tamaño de 8.669 pb y tiene ocho exones. En frijol Lima, este gen se encuentra ubicado en la cadena negativa del contig-1891 del ensamblaje producido en la presente investigación, con un porcentaje de identidad del 86 % con respecto a frijol común, un tamaño de 7.801 pb, y ocho exones. En la figura **5-1** se detallan los tamaños de cada uno de los exones e intrones en estos genes. En esta gráfica se puede observar una alta conservación en la estructura y tamaño del gen SHP1, en las tres especies evaluadas (ver tabla **5-2**).

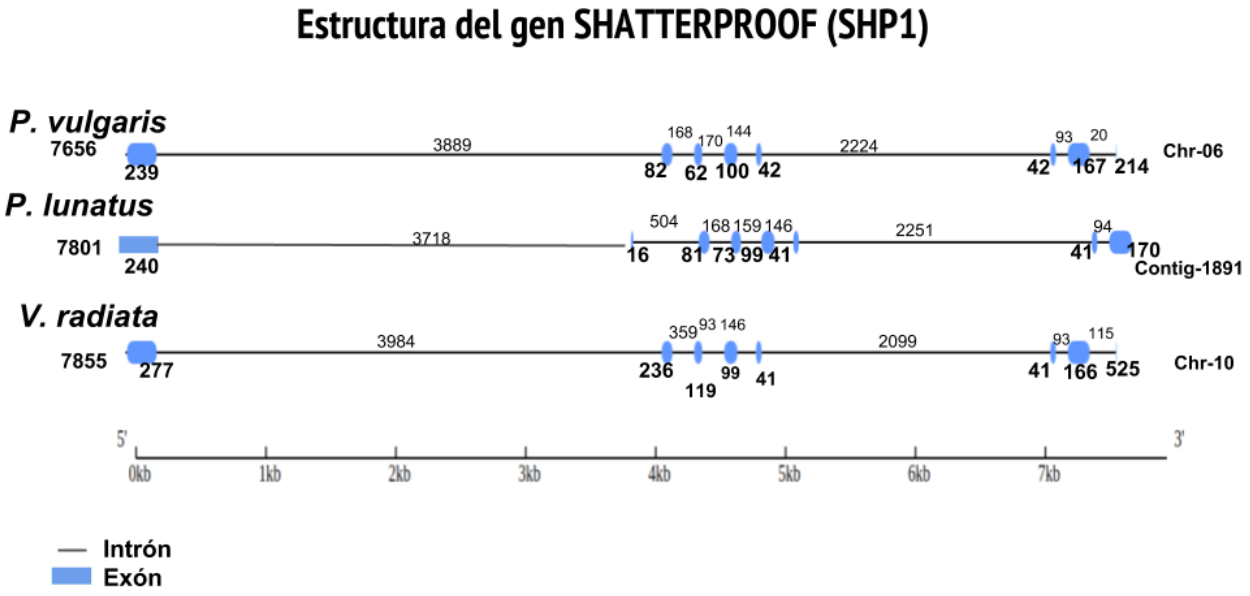


Figura 5-1.: Estructura del gen *SHP1* en frijol común, frijol Lima y frijol mungo.

Al evaluar la región genómica del gen *SHP1* (fig **5-2**) utilizando la herramienta Mauve, se observaron dos grandes bloques colineales, uno marcado en color rojo (longitud: 56.252 pb, peso: 141.571) y otro marcado en color verde (longitud: 58.568 pb, peso: 53.050), y otro bloque colineal más pequeño marcado en amarillo (longitud:1.061 pb, peso:3.560). Estos bloques colineales se detectaron con un peso mínimo de 3.560 y una longitud de 1.061 pb. Las regiones grises al interior de los bloques colineales significan aquellas secuencias que son específicas para cada especie y por lo tanto no compartidas. Con relación a esto, se observa que el frijol mungo posee un mayor número de regiones específicas de su linaje evolutivo, lo cual está de acuerdo con el hecho que esta especie pertenece a un género diferente (*Vigna*) al de las otras dos especies de frijol (*Phaseolus*). Es importante destacar que en la región evaluada, solamente para frijol mungo se observa que el bloque con color amarillo está invertido, mientras que los restantes dos bloques (verde y rojo) se conservan en las tres especies.

Con respecto al contenido génico de la región evaluada (fig **5-3**), el genoma de frijol común incluye 24 genes, de los cuales nueve se encuentran en la cadena negativa. En frijol Lima se identificaron 22 genes de los cuales seis se ubican en la cadena negativa, todos éstos en una región especifíca que comprende del gen 12-Pl al gen 17-Pl. En frijol mungo se ubicaron 12 genes, siete de éstos en la cadena negativa. Al considerar a los genes 12-Pv, 13-Pl y 4-Vr como posibles copias ortólogas del gen *SHP1*, se observa que éstos comparten el mismo sentido. En las tablas C-1, C-2 y C3 (anexo C), se detalla el número de genes en cada sub-región. Se puede observar que frijol común y frijol Lima comparten el mayor número de genes en esta región subgenómica que con frijol mungo, lo cual está relacionado con la distancia evolutiva entre estas especies.

Se identificaron 3 genes ortólogos adicionales en esta subregión genómica (genes A, B y C), representados en la gráfica **5-1**, con color verde. Los genes A, B y C conservan el mismo orden en las tres especies. El gen A se identificó con un porcentaje de identidad mínimo del 88 %, donde este gen en frijol común y frijol Lima presenta un tamaño similar de su secuencia con 2.396 pb y 2.621 pb, respectivamente, y en ambas especies tiene 5 exones. En cuanto al gen A en frijol mungo, su tamaño es de 1064 pb en 3 exones. El gen B presentó un porcentaje de identidad del 86 %, en las tres especies conserva la orientación negativa y un tamaño similar: en frijol común con 7.594 pb, en frijol Lima con 7.273 pb y en frijol mungo con 7.893 pb, pero varía en el número de exones:12, 13 y 19, respectivamente. Con un porcentaje de identidad del 85 % se identificó el gen C, el cual cuenta en frijol común y frijol Lima con seis exones y un tamaño similar de 4.554 pb y 4.422 pb, respectivamente. En el caso del frijol mungo, el gen C presenta el mismo número de exones pero con un tamaño menor en la secuencia de 2.903 pb. En general, el conjunto de genes ortólogos que acompañan en la misma región genómica al gen *SHP1*, tienen una identidad y estructura similares entre las especies comparadas.

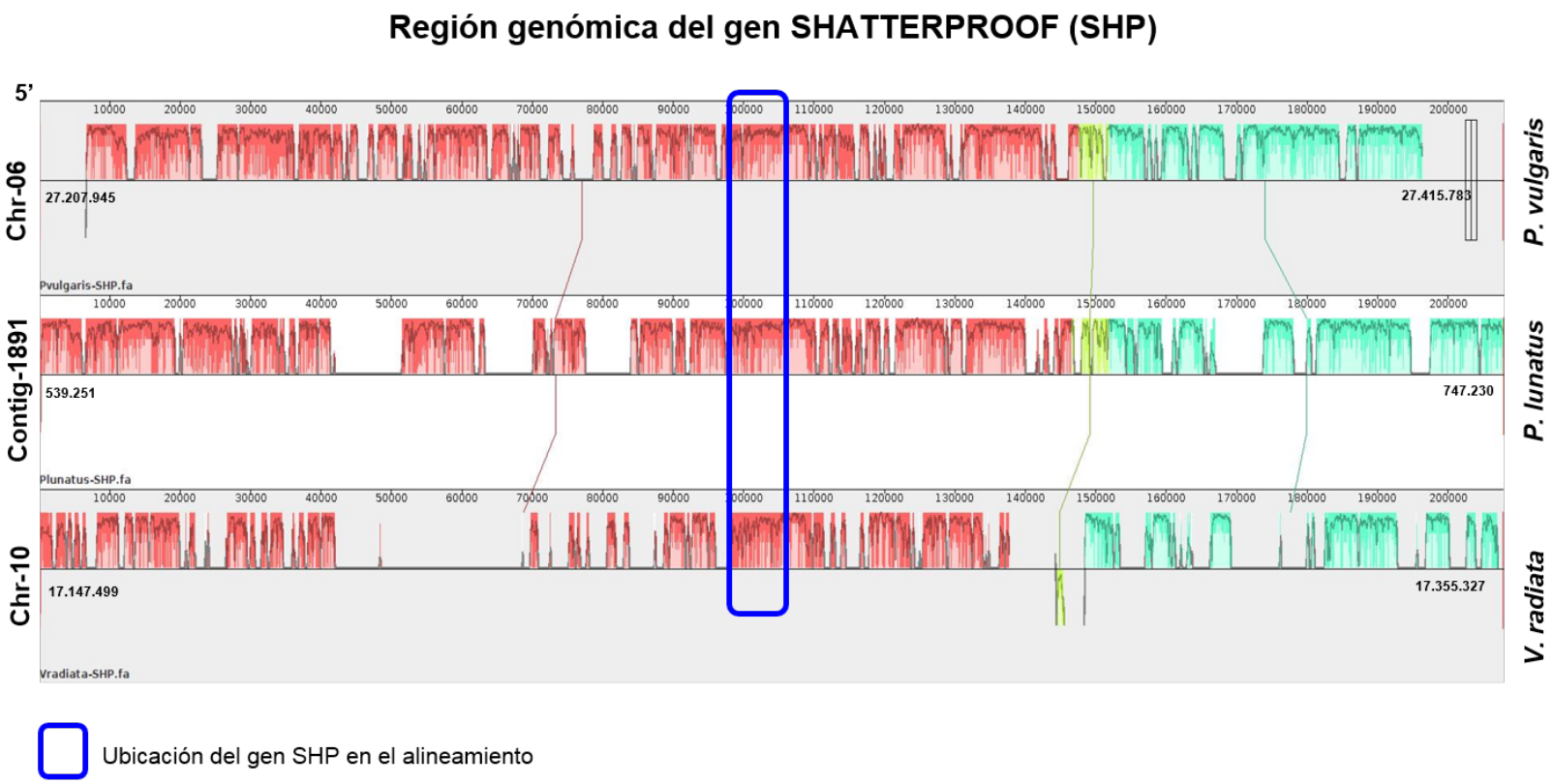


Figura 5-2.: Caracterización de la región subgenómica del gen *SHP1* para frijol común, frijol Lima y frijol mungo

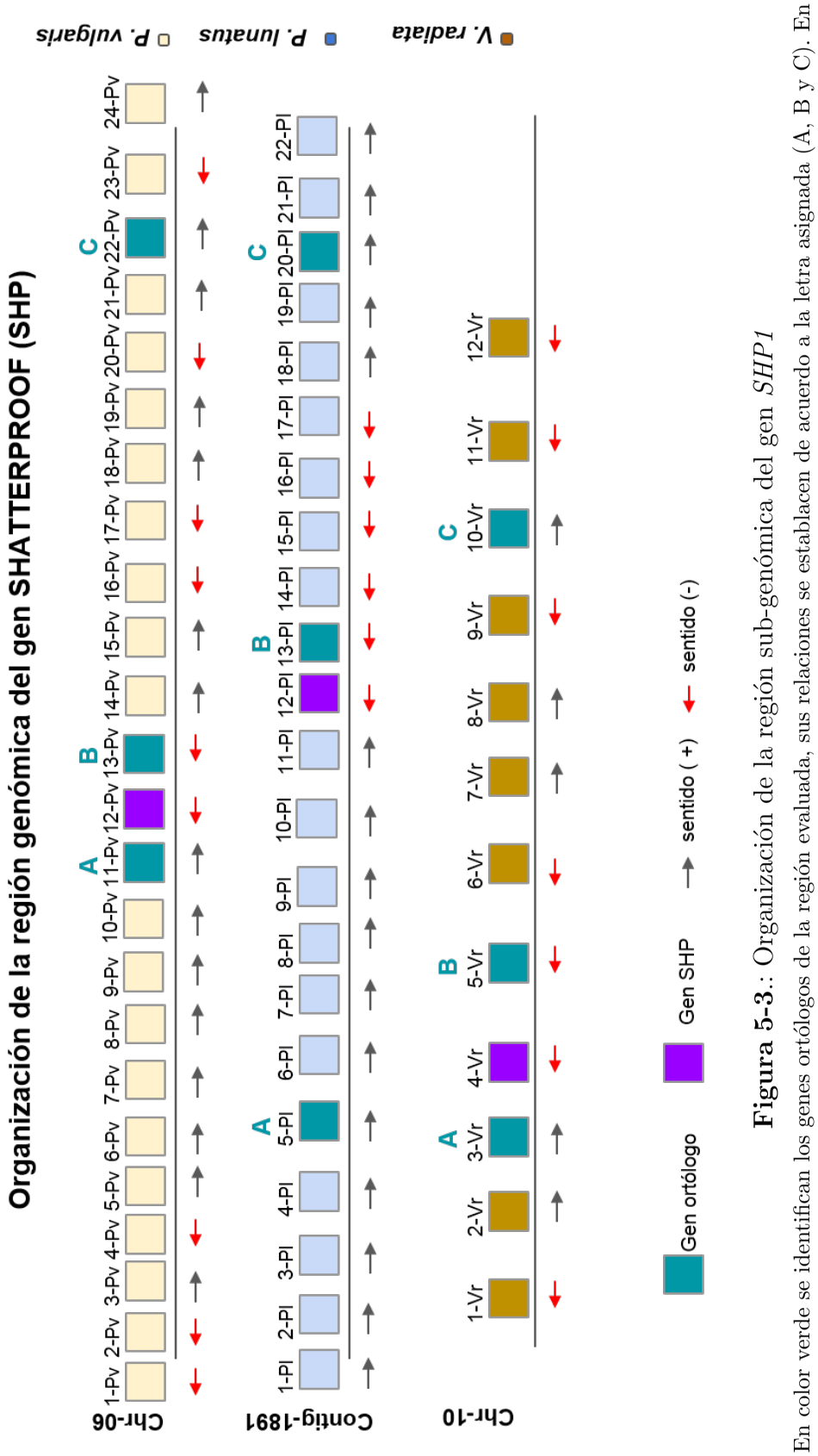


Figura 5-3.: Organización de la región sub-genómica del gen *SHP1*

En color verde se identifican los genes ortólogos de la región evaluada, sus relaciones se establecen de acuerdo a la letra asignada (A, B y C). En color morado se representa el gen candidato de *SHP1*.

5.3.3. Caracterización de la región subgenómica asociadas al gen INDEHISCENT (IND)

INDEHISCENT(*IND*) está relacionado con una proteína atípica con dominio basic helix-loop-helix (bHLH), que incide en la diferenciación de las células en la zona de dehiscencia de la silicua en *Arabidopsis*. En frijol común, este gen (id Phvul.002G0271000) presenta un tamaño de 1.340 pb con un exón de 846 pb ubicándose en el cromosoma 2 en la cadena positiva (fig **5-5**). En frijol mungo, el homólogo de este gen (Vradi01g11070) se identificó con un porcentaje de identidad del 93 % con respecto a frijol común, con una anotación funcional de factor de transcripción con el dominio bHLH, presenta un tamaño de 842 pb, en la cadena positiva. En frijol Lima se identificó un posible candidato de este gen (2-Pl) con un exón y una longitud total de 843 pb. Se debe destacar que este gen conserva su estructura con un único exón como se observa en la figura **5-4** en las tres especies evaluadas.

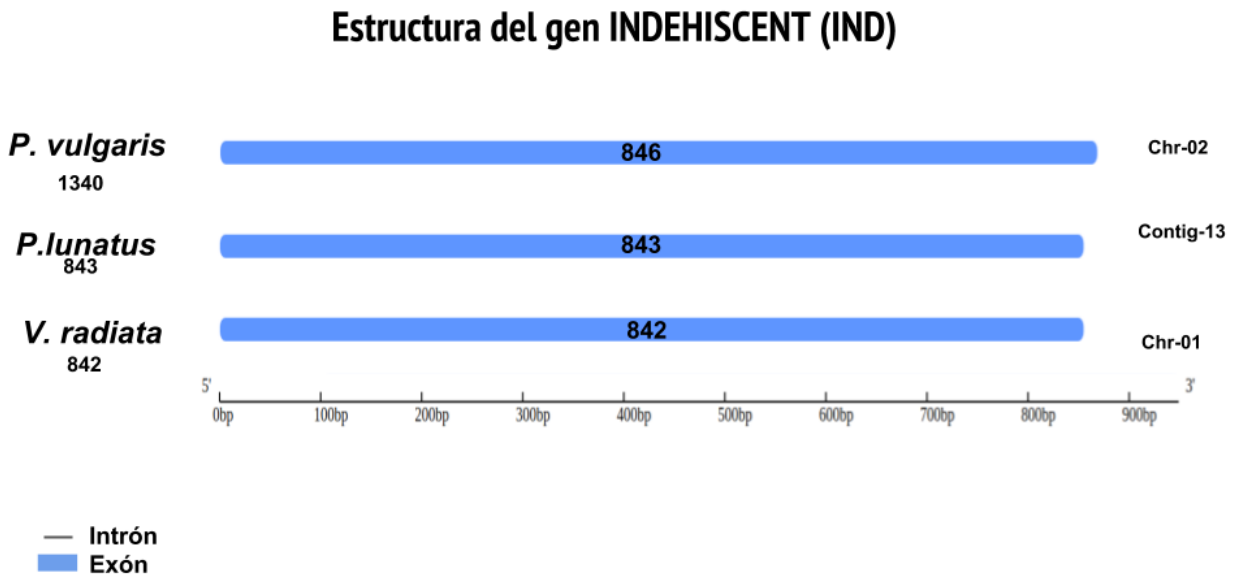


Figura 5-4.: Estructura del gen *IND* en frijol común, frijol Lima y frijol mungo

A nivel estructural, en la sub-región genómica evaluada (fig **5-5**) se observaron siete bloques colineales identificados con un peso mínimo de 1.254 y una longitud de 1.917 pb. Los principales bloques se obervan en color amarrillo (longitud: 159.676 pb, peso: 72.800), color rojo (longitud: 60.883 pb, peso: 41.650) y color verde (longitud:7.715 pb, peso:12.473). El bloque de color verde al final de la figura evidencia una inversión en frijol mungo con relación a frijol común y frijol lima.

En cuanto al bloque que contiene el gen *IND* (color morado), presenta una longitud 4.242 pb y se identificó con un peso de 4.749. En este bloque es evidente que frijol Lima en esta región ha experimento un rearreglo cromosómico con relación al frijol común. El hecho que existan inversiones en esta región entre frijol común y frijol Lima conlleva implicaciones prácticas, debido a que en el cromosoma 2 de frijol común se ha identificado por medio de mapeo de QTL (Quantitative Trait Loci) un QTL llamado *St* asociado a dehiscencia de la vaina y muy cercano al gen *IND* [66, 49]. Los rearreglos observados en esta región están de acuerdo con estudios de citogenética comparativa entre frijol común y frijol Lima que han reportado que aunque existe una significativa conservación de sintenia entre las especies, para el cromosoma 2 se reportó cambio en la posición del centrómero, posiblemente como consecuencia de una inversión pericentromérica [10].

En cuanto al contenido génico de de la región genómica que contiene el gen *IND* (fig **5-6**), se observa que frijol común cuenta con 17 genes, de los cuales cinco están en la cadena negativa, en frijol lima se identificaron solamente 6 genes, uno de ellos en la cadena negativa. Con respecto a la región de frijol mungo, esta cuenta con 11 genes, tres en la cadena negativa. En las tablas C-4 y C-5 (Anexo-C), se encuentra mayor información sobre los genes de esta región. Es importate destacar que en la zona genómica de interés no se hallaron genes ortólogos adionales como se identififcaron los tres genes colineales al gen *SHP*.

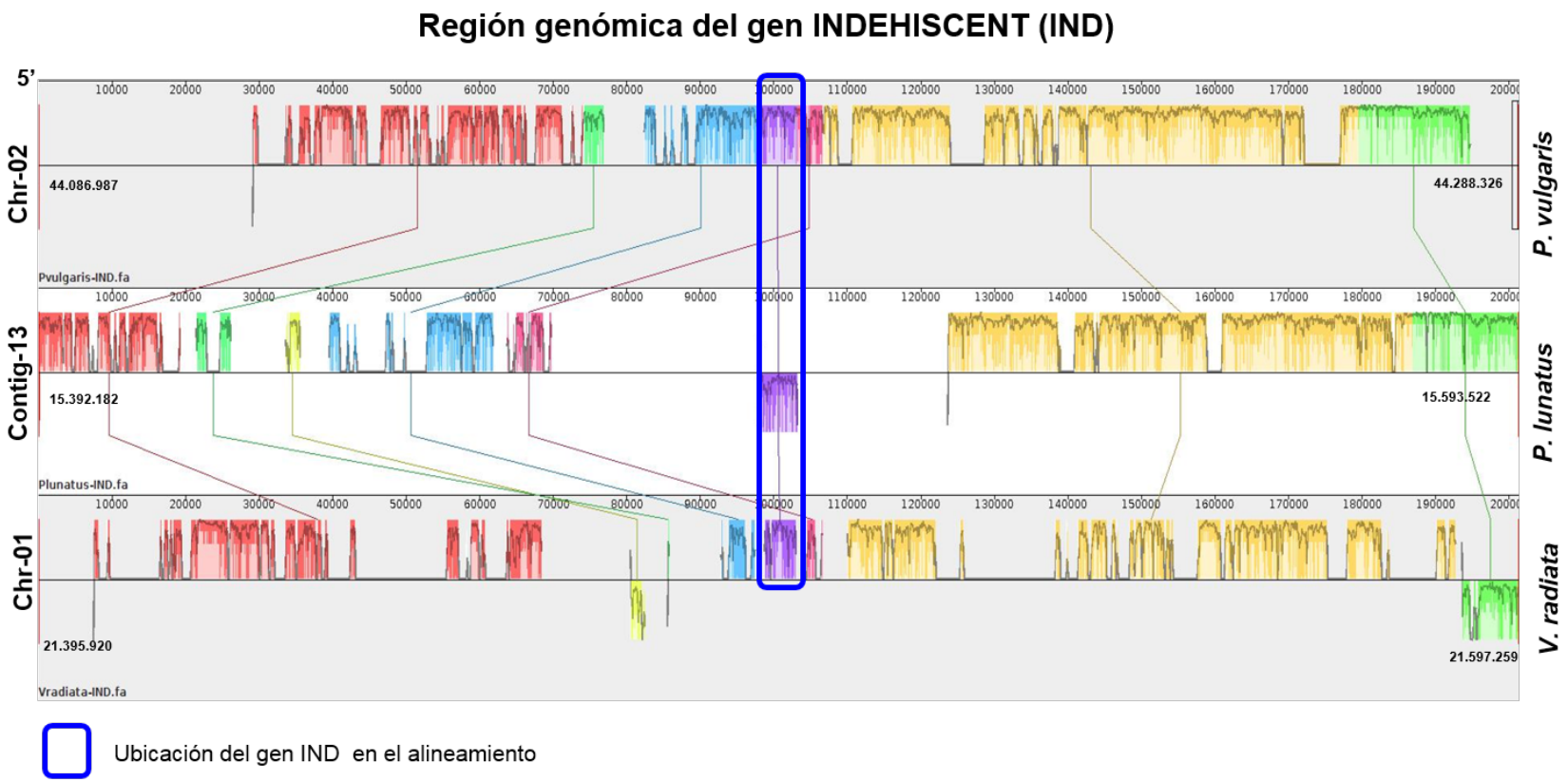


Figura 5-5.: Caracterización de la región subgenómica del gen IND para frijol común, frijol Lima y frijol mungo

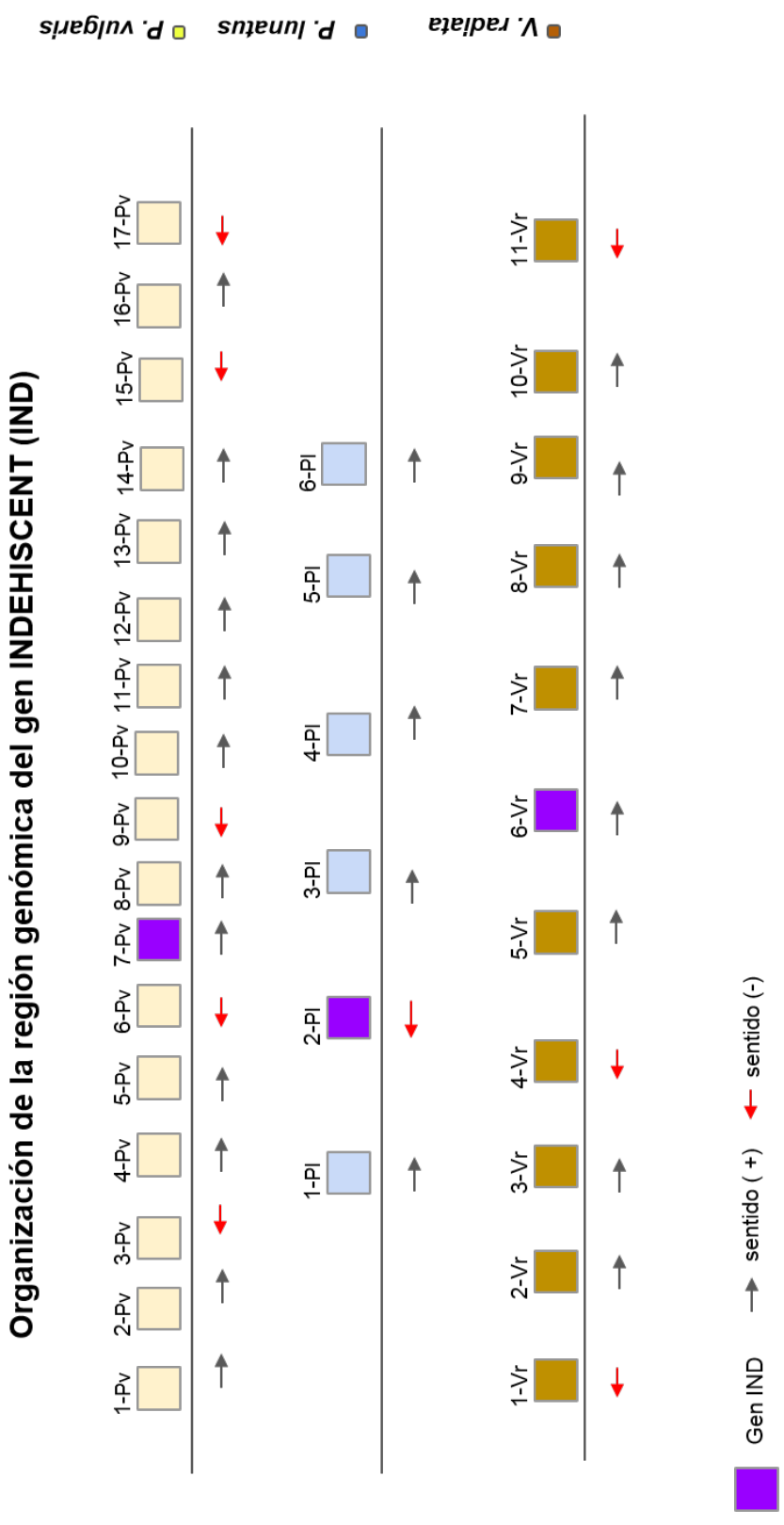


Figura 5-6.: Organización de la región sub-genómica del gen *IND*

5.3.4. Caracterización de la región subgenómica asociadas al gen ALCATRAZ (ALC)

La función del gen *ALCATRAZ* (*ALC*) ha sido identificada en relación con la familia de factores de transcripción myc/bHLH, incidiendo en el sitio de separación de la valva durante la dehiscencia del fruto en *Arabidopsis* [73]. *ALC* en frijol común (ID Phvul.001G023200) se encuentra ubicado en el cromosoma 1 en la cadena negativa (fig **5-8**), tiene un tamaño de 3.723 pb con seis exones (fig **5-7**), presenta dos isoformas que se diferencian en el tamaño del segundo exón (90pb para la isoforma uno y 87 para la isoforma dos).

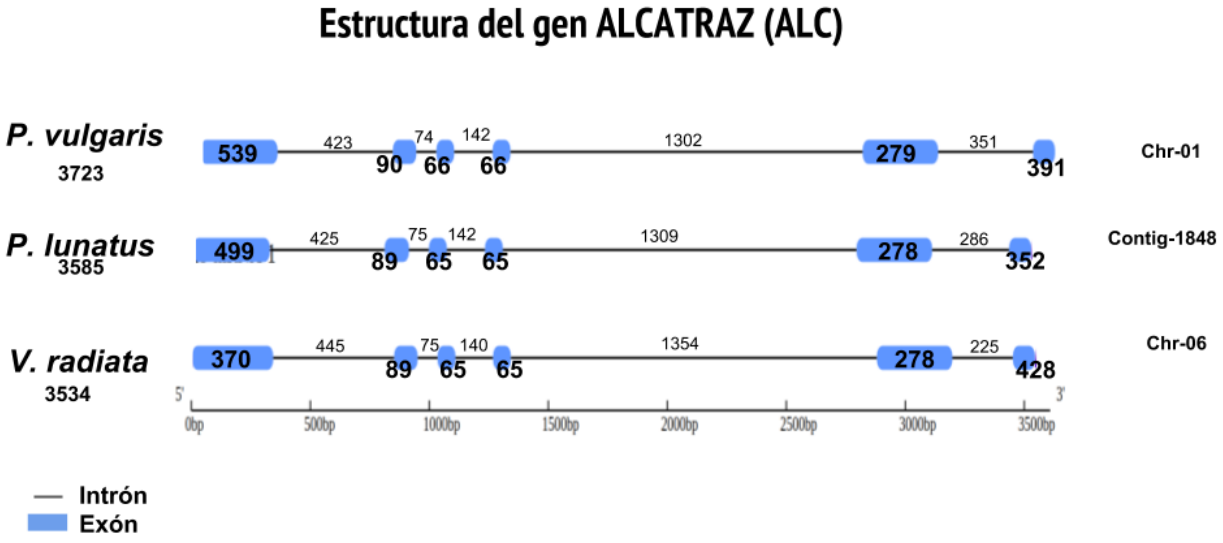


Figura 5-7.: Estructura del gen ALC en frijol común, frijol Lima y frijol mungo.

En frijol mungo, un posible gen ortólogo de *ALC* se encuentra en el cromosoma 6 (ID Vradi06g15060), con un porcentaje de identidad del 88 % con respecto a frijol común. Este gen candidato en frijol mungo se compone de 3.534 pb, con una arquitectura de seís exones (fig **5-7**). Su anotación funcional [44] indica su acción como factor de transcripción con un motivo basic helix-loop-helix (bHLH), al igual que el reportado en *Arabidopsis* [73]. En frijol Lima se identificó este gen en el presente estudio con un porcentaje de identidad del 92 % en el contig 1.448 con seis exones y un tamaño de 3.585 pb. En la fig **5-7**, se puede identificar que, en relación al tamaño total del gen y número de exones, frijol Lima y frijol común son muy similares. Al comparar a nivel estructural la región genómica que contiene el gen *ALC*(fig **5-8**), se observa que esta región consiste de un solo bloque colineal (longitud: 202.539 pb,

peso: 197.588) que por definición no contiene rearreglos cromosómicos, lo que indica un alto grado de conservación para esta región sub-genómica.

En cuanto al contenido génico de la región evaluada que contiene el gen *ALC* (**5-9**), la secuencia genómica de frijol común cuenta con 20 genes, 14 están en la cadena negativa. En frijol lima se observan 21 genes, de los cuales sólo 4 estan en la cadena positiva. En frijol mungo esta región cuenta con 15 genes, con 3 de ellos en la cadena positiva. El tamaño de los genes, el número de exones e introns para cada una de las especies de frijol, se reportan en las tablas **C-6**, **C-7** y **C-8** (Anexo C).

En esta región se identificaron 5 genes ortólogos cercanos al gen de interés. El gen ortólogo identificado con la letra A, presentó un 91 % de identidad, este gen varía considerablemente en tamaño de de la secuencia y número de exones entre frijol común (2.077 pb , 2 exones), frijol Lima (1.220 pb , 1 exones) y frijol mungo (5.324 pb, 5 exones). Con un 90 % de identidad, el gen B presenta un tamaño muy similar de 5.384 pb en frijol común, 5.063 pb en frijol Lima y 5.047 pb en frijol mungo, en las tres especies presenta 6 exones. El gen C presenta en frijol común y frijol mungo con un tamaño 3.537 pb (4 exones) y 3.453 pb (9 exones), respectivamente, y mientras que en frijol Lima tiene un tamaño inferior (1.224 pb) con respecto a frijol mungo y frijol común, pero con 2 exones. El gen D, identificado con el 93 % de identidad, presenta en las tres especies 9 exones. En el caso de Frijol Lima y frijol común presentan un tamaño similar de secuencia, 5.005 pb y 5.080 pb, mientras que para frijol mungo su tamaño de secuencia es de 4.413 pb. Con el 88 % de identidad, el gen E se identificó en frijol común con 2.014 pb y 3 exones, en frijol Lima con 2.539 pb y 3 exones, y en frijol mungo con 1.899 pb y 4 exones. En esta sub-región genómica se observa que el gen de interés está flanqueado por un gen ortólogo (gen A) al lado izquierdo de acuerdo a la representación de la figura **5-9** y cuatro genes ortólogos al lado derecho. Adicionalmente al comparar el anterior conjunto de genes ortólogos encontrados en la región del gen *ALC*, con los del gen *SHP1*, se evidencia una mayor conservación en los genes que colindan la región del gen *SHP1* debido a la menor variación que presentan en su tamaño y número de exones.

[H]

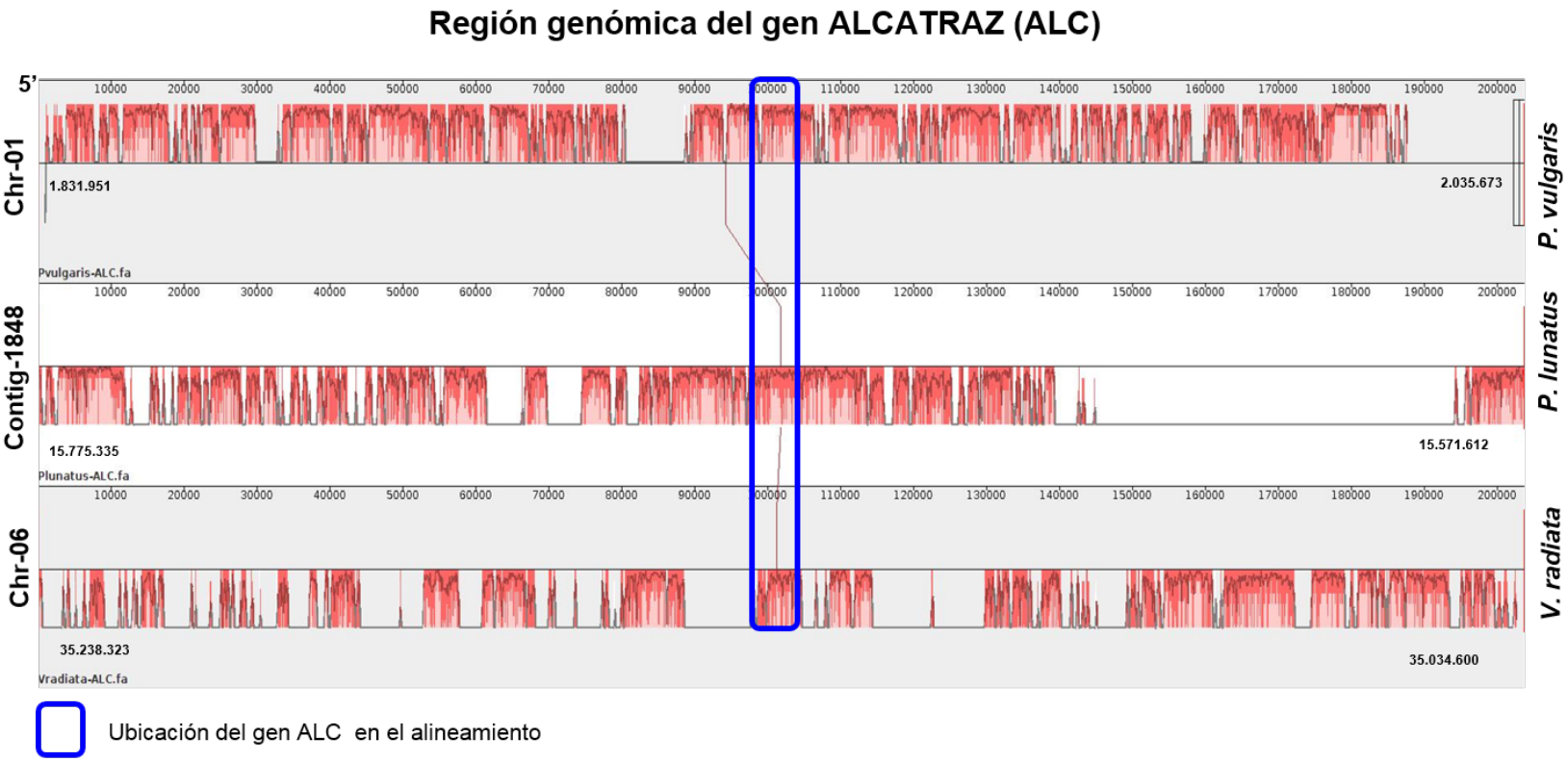


Figura 5-8.: Caracterización de la región subgenómica del gen ALC para frijol común, frijol Lima y frijol mungo

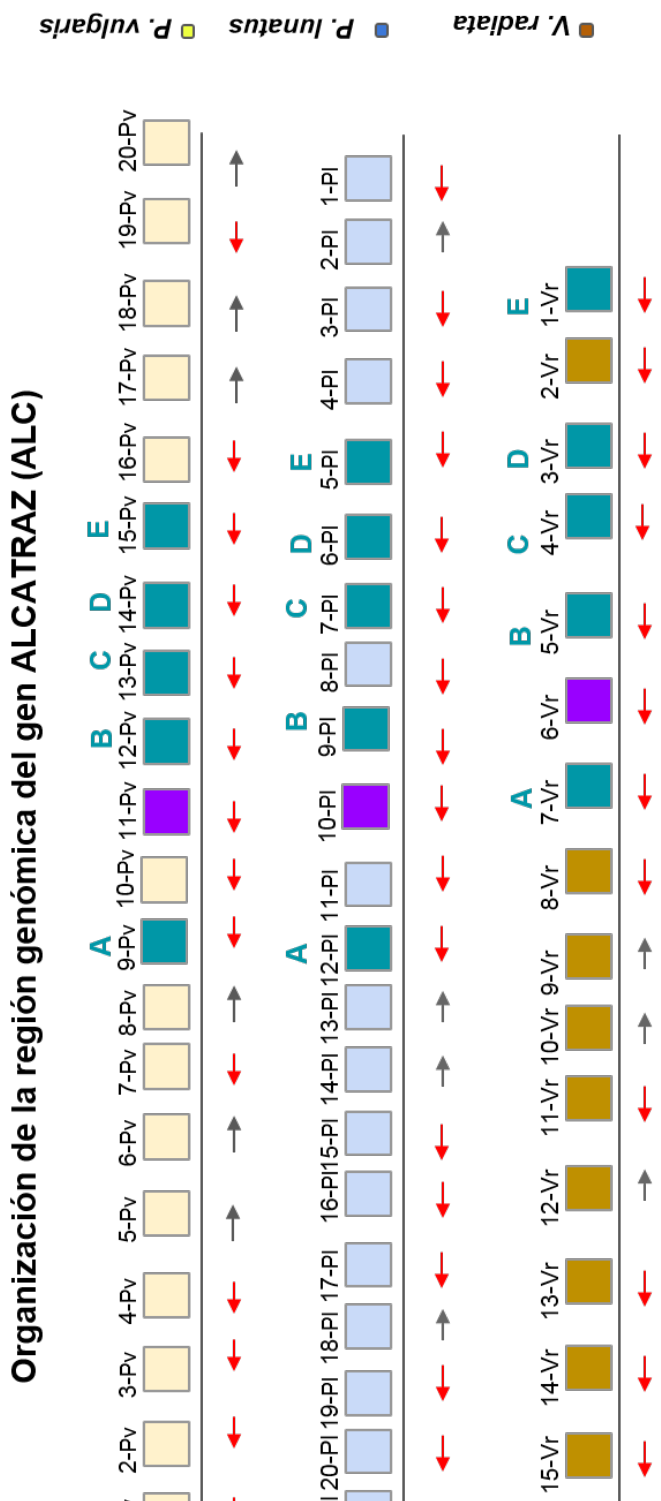


Figura 5-9.: Organización de la región sub-genómica del gen ALC

En color verde se identifican los genes ortólogos de la región evaluada, sus relaciones se establecen de acuerdo a la letra asignada (A, B , C, D y E). En color morado se representa el gen candidato de *ALC*.

5.3.5. Caracterización de la región subgenómica asociadas al gen FRUITFULL (FUL)

FRUITFULL (*FUL*) es un gen tipo MADS-box, involucrado en la expansión y diferenciación de las valvas del fruto en *Arabidopsis*. En frijol común (id Phvul.009G203400) presenta una longitud de 9.086 pb, con ocho exones y siete intrones, dos de ellos de gran tamaño (3.394 bp y 2.565 bp). En frijol mungo se identificó un gen candidato para *FUL* (Vradi02g13700) con un tamaño de 7.934 pb con 8 exones, su función está asociada al grupo MADS-box [42]. En frijol lima se encontró un homólogo de *FUL* en el contig 1.846, con 8.892 pb y 8 exones. En la figura **5-10**, se observa que frijol lima y frijol común presentan una alta similitud en cuanto a los tamaños de los exones e intrones, sugiriendo al gen identificado en frijol Lima como un posible ortólogo de frijol común.

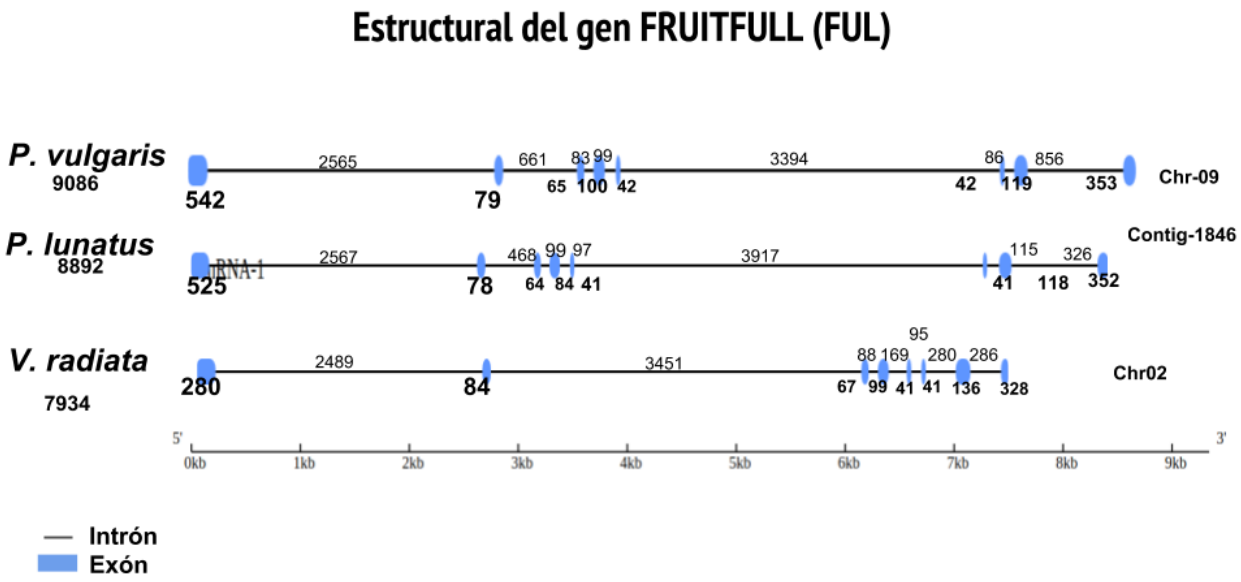


Figura 5-10.: Estructura del gen *FUL* en frijol común, frijol Lima y frijol mungo

En la región sub-genómica que contiene el gen *FUL* (fig **5-11**) se puede observar que existen cuatro bloques colineales, identificados con un peso minimo de 6.262, sin re-arreglos cromosómicos entre ellos a excepción del pequeño bloque de color azul que aparece en orden invertido entre frijol común y frijol mungo. En el bloque colineal verde que contiene el gen *FUL* (peso: 89.882) se corrobora un alto grado de homología en la mayoria de la región evaluada.

En cuanto al contenido génico de esta región, frijol Lima reporta la mayor cantidad de genes (21 genes), seguido de frijol común con 15 y frijol mungo con 9 (figura **5-12**). En las tablas C-9, C-10 y C-11 (Anexo C) se encuentra una descripción de cada uno de estos genes en cuanto a su longitud, número de exones e intrones.

En esta sub-región se hallaron seis genes ortólogos (figura **5-12**). Los genes ortólogos marcados con la letra A se identificaron con el 92 % de identidad. En las tres especies este gen presenta un único exón, en frijol mungo y frijol Lima reportan un tamaño similar alredor de las 1.300 pb, mientras que en frijol común contiene 1.000 pb adicionales. Con el 91 % de identidad se identificó el gen B, en frijol Lima y en frijol común conservan el mismo número de exones (4), y un tamaño de 2.283 pb y 2.026 pb respectivamente, mientras que en frijol mungo su tamaño es menor con 1.339 pb y 3 exones. El gen C presentó una identidad del 94 %. Este gen en frijol mungo se diferencia de frijol Lima y frijol común en más de 6.000 pb y 20 exones. frijol Lima y frijol común presentan el mismo número de exones (4) y un tamaño de 2.283 pb y 2.026 pb, respectivamente. El gen D se identificó con 88 % de identidad, donde en frijol Lima y frijol común reporta un tamaño similar de 12.541 pb y 12.785 pb, respectivamente, sin embargo se diferencian en el número de exones (8 y 11, respectivamente). El gen E se identificó con un 85 % de identidad donde en frijol mungo y frijol Lima cuenta con 3 exones y un tamaño aproximado de 1.600 pb, mientras que en frijol común su tamaño es el doble al igual que el número de exones. Finalmente, con un porcentaje de identidad del 88 % se identificó el gen F, el cual es muy similar en tamaño (14.333, 11.565 y 11.996) y número de exones (13, 11 y 10) en frijol común, frijol Lima y frijol mungo, respectivamente.

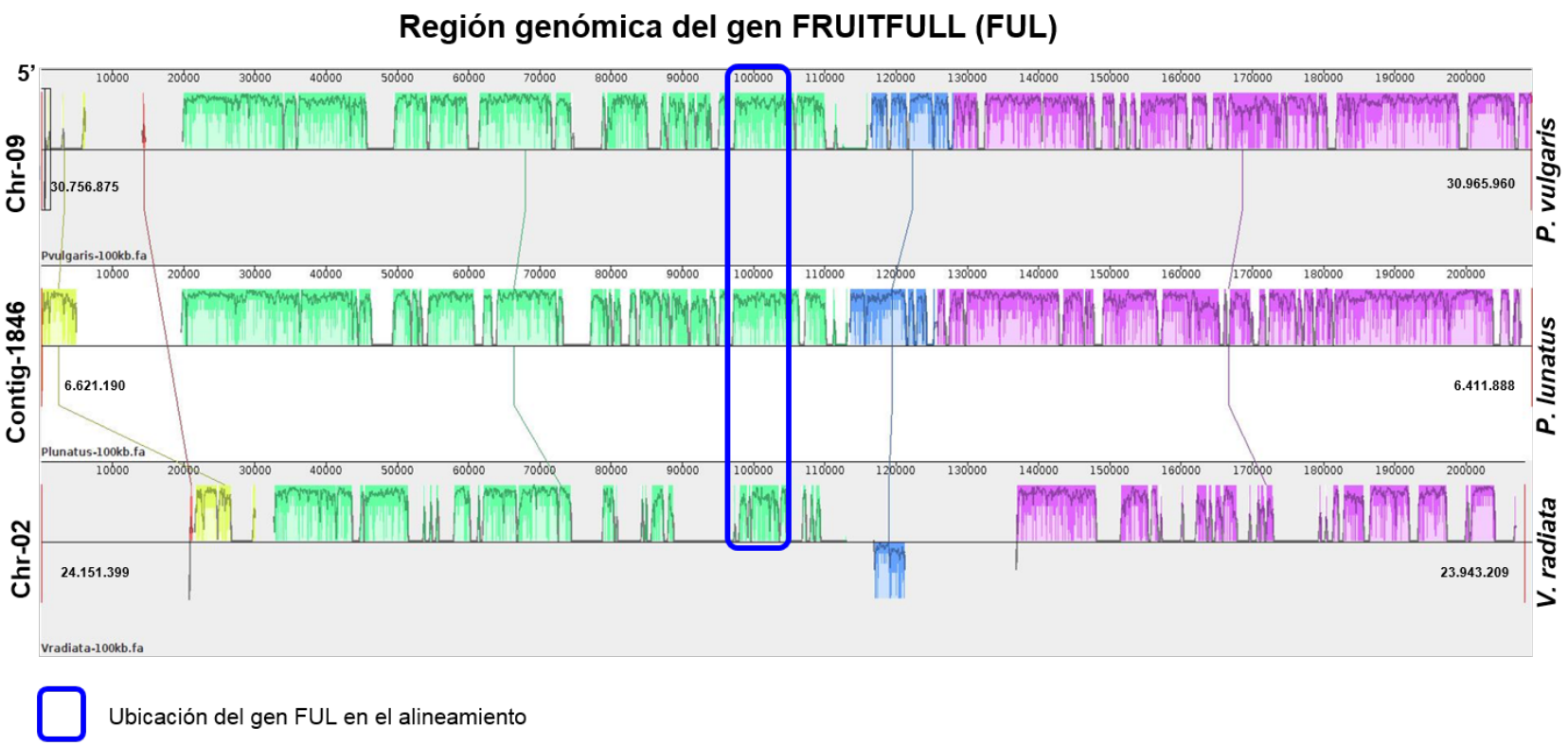


Figura 5-11.: Caracterización de la región subgenómica del gen FUL para frijol común, frijol Lima y frijol mungo

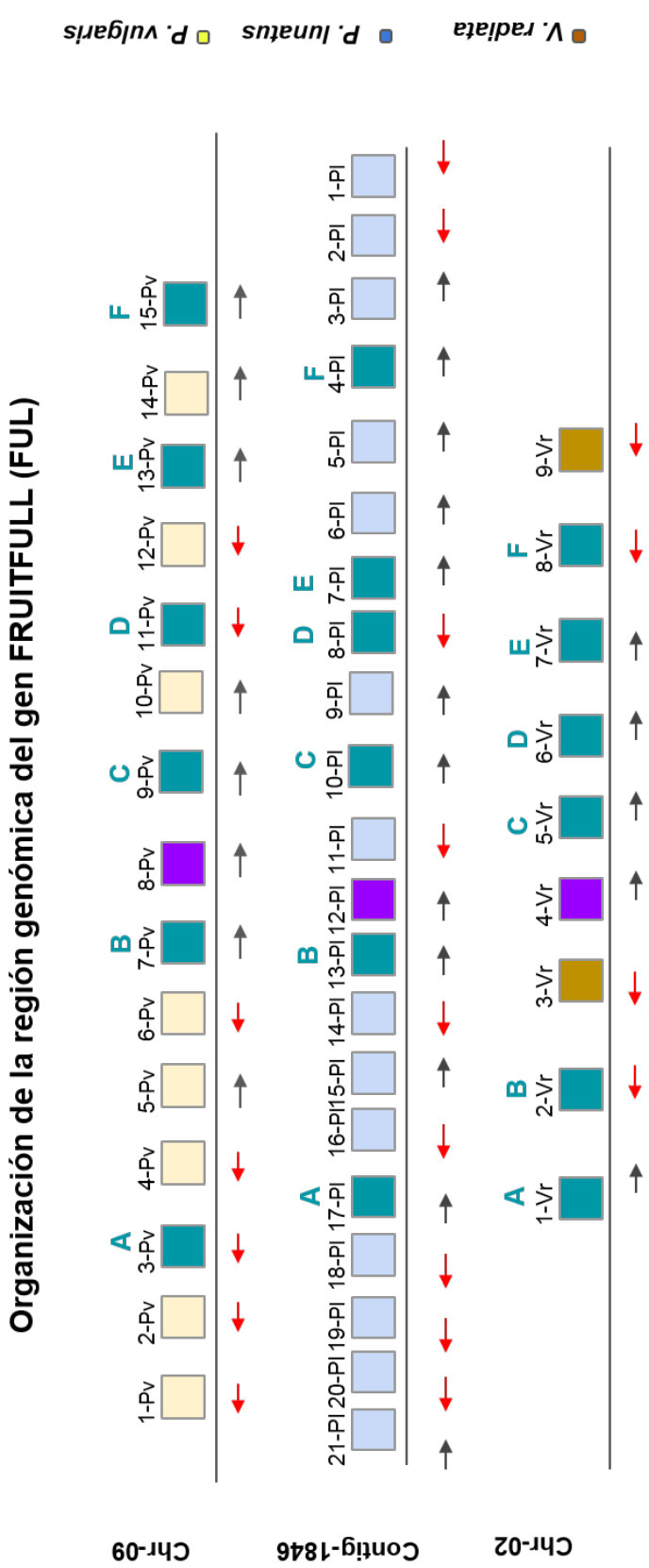


Figura 5-12.: Organización de la región sub-genómica del gen FUL

En color verde se identifican los genes ortólogos de la región evaluada, sus relaciones se establecen de acuerdo a la letra asignada (A, B, C, D, E y F). En color morado se representa el gen candidato de *FUL*.

5.3.6. Caracterización de la región subgenómica asociada al gen NST1 (NAC SECONDARY WALL THICKENING PROMOTING FACTOR 1)

El gen *NST1* (*NAC SECONDARY WALL THICKENING PROMOTING FACTOR 1*) se ubica en el cromosoma 10 (id Phvul.010G118700) de frijol común, con un tamaño de 2.152 pb con 3 exones, en la cadena negativa. En el frijol mungo se identificó el gen *NST1* con un porcentaje de identidad del 92 % con relación a frijol común, éste presenta una longitud de 2.111 pb con tres exones. En frijol lima, se identificó un homólogo del gen *NST1*, con una longitud de 2.437 pb, con 3 exones. Al comparar los tres genes (fig **5-13**), se observa que frijol Lima y frijol común conservan una alta similitud, en cuanto al tamaño total del gen, número y tamaño de los exones e intrones, excepto por el último exón que en frijol Lima tiene 222 pb de bases adicionales.

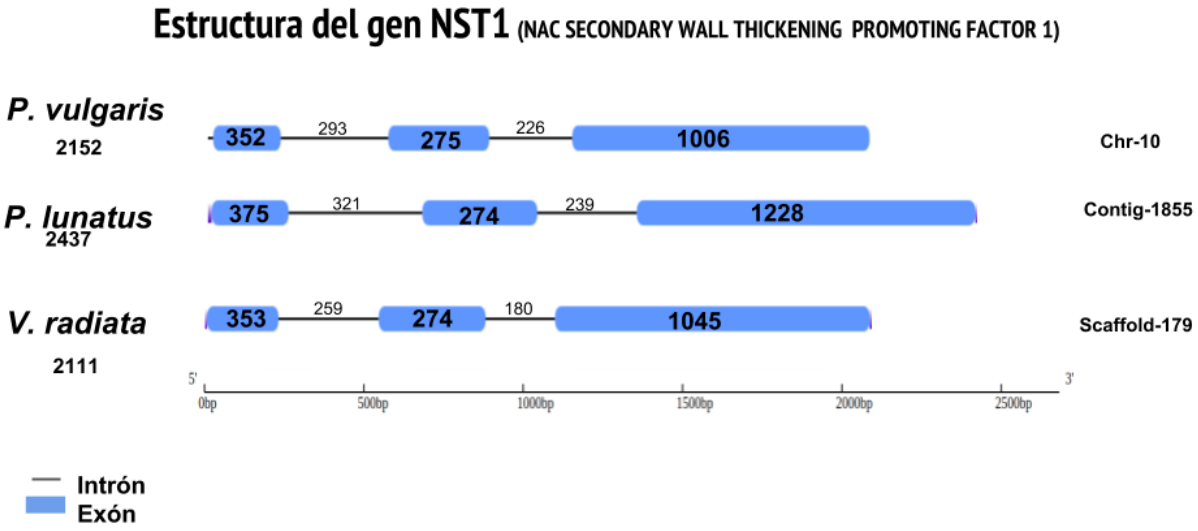


Figura 5-13.: Estructura del gen *NST1* en frijol común, frijol Lima y frijol mungo

A nivel estructural, en la región subgenómica que contiene el gen *NST1* (**5-14**) se identificaron con un peso mínimo de 4.142, seis bloques colineales entre las tres especies comparadas, y un solo bloque compartido adicional entre frijol mungo y frijol lima identificado con color azul agua marina, el cual está en orientación del complemento reverso con relación a la orientación en frijol lima. Otro importante evento que se observa en la gráfica **5-14** es que el bloque colineal de color azul se comparte solo entre frijol común y frijol mungo, sin estar

presente en frijol Lima. En cuanto al bloque donde se ubica el gen *NST1*, (color morado, peso:93.110), se observa que está en el mismo sentido con en las tres especies de frijol.

En cuanto al contenido génico en la región que contiene el gen *NST1* (fig **5-15**), en el frijol común se reportan 10 genes, de éstos uno solamente en la cadena positiva. En el frijol lima se reportan 18 genes, con los genes del 6-Pl al 13-Pl en la cadena negativa, entre ellos el gen candidato *NST1* (gen 10-Pl). En el frijol mungo se reportan 14 genes, de éstos solamente tres se encuentran en la cadena positiva. En las tablas C-12, C-13 y C-14 (Anexo C) se encuentra una descripción de cada uno de estos genes en cuanto a su longitud, número de exones e introns. Adicionalmente en esta región se identificaron 3 genes ortológos (A, B y C) (color verde en la fig **5-14**). Es importante anotar que el gen *NST1* se encuentra entre los genes B y C en frijol común y frijol Lima, mientras que en frijol mungo está por fuera de este rango, lo que sugiere un evento de inversión. El gen A se identificó con una identidad del 88 %, presenta un tamaño similar en las 3 especies, en frijol común con (5.425 pb, en frijol Lima con 6.219 pb y en frijol mungo con 4.447 pb, con 4 exones en frijol común y 5 exones en las dos especies restantes. Con una identidad del 92 % se identificaron los genes B y C. En frijol común el gen B tiene 7 exones y un tamaño total de 2.672 pb, mientras que en frijol Lima y frijol mungo el gen B comparte el mismo número de exones de 5 y un tamaño alredor de 2.600 pb. En los genes identificados con la letra C todos cuentan con un único exón, y tamaños diferentes, 1.157 pb (frijol común), 539(frijol Lima) y 989 (frijol mungo).

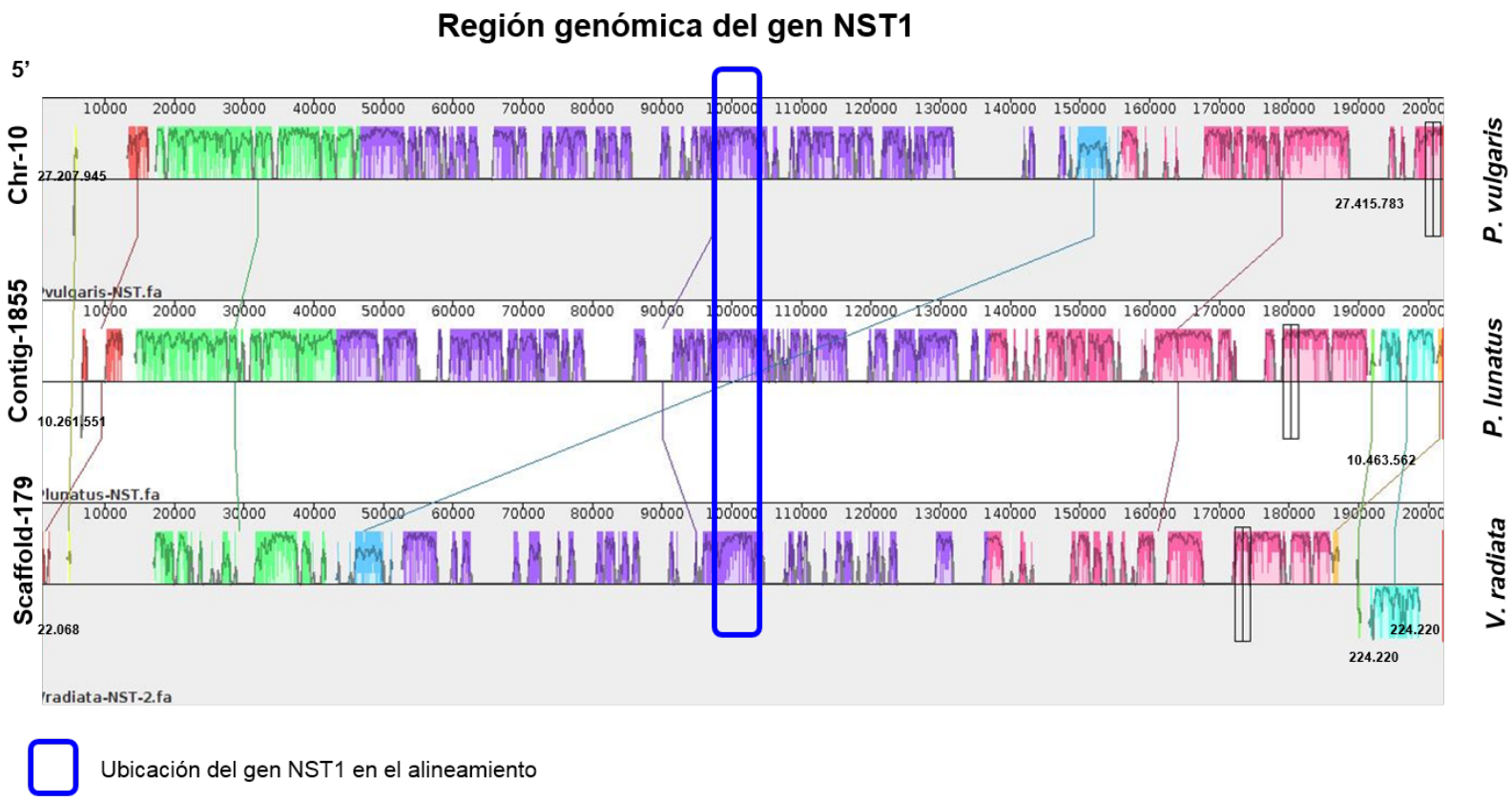


Figura 5-14.: Caracterización de la región subgenómica del gen *NST-1* para frijol común, frijol Lima y frijol mungo

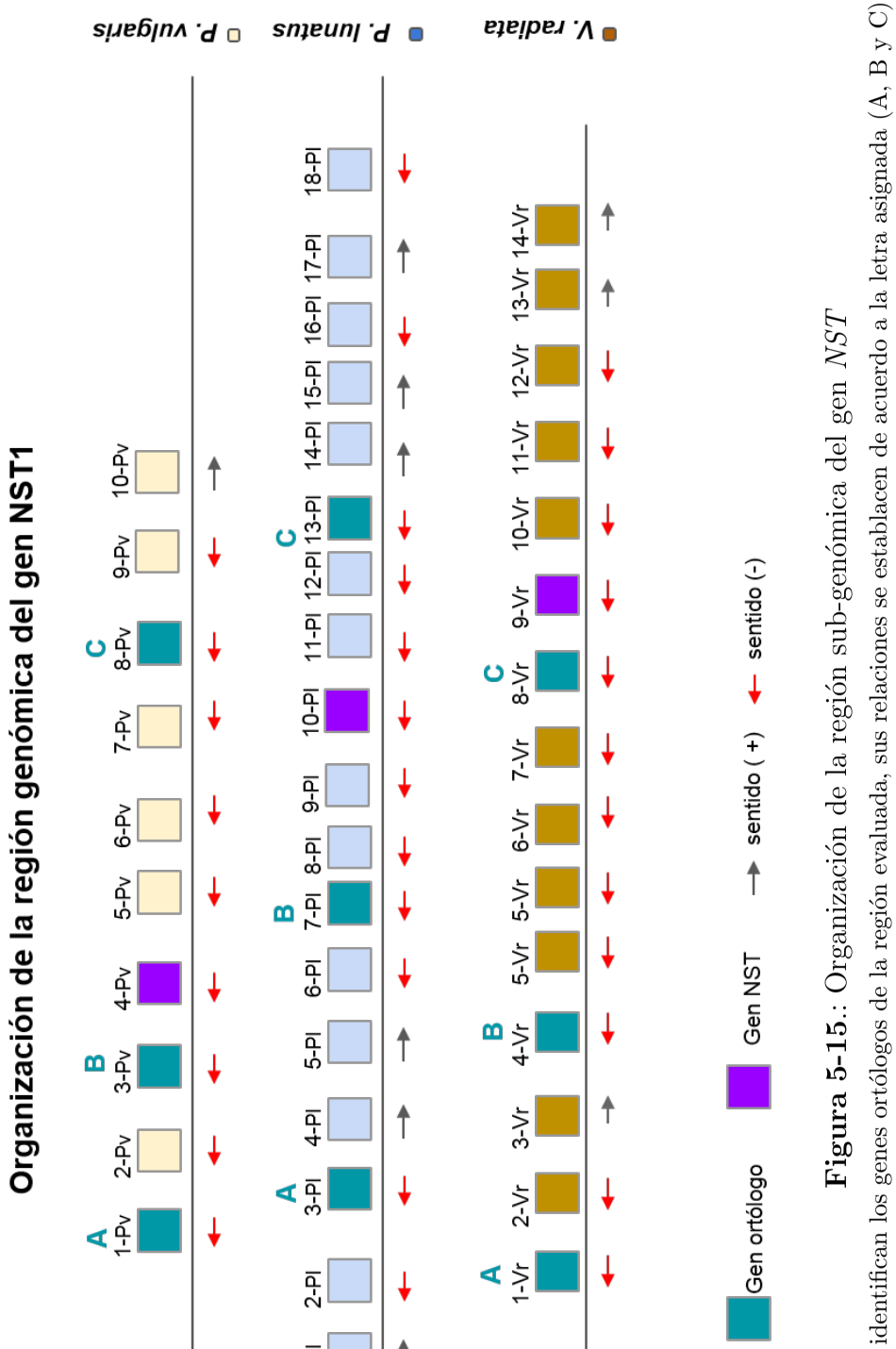


Figura 5-15.: Organización de la región sub-genómica del gen *NST*
En color verde se identifican los genes ortólogos de la región evaluada, sus relaciones se establecen de acuerdo a la letra asignada (A, B y C). En color morado se representa el gen candidato de *NST*.

Finalmente, en la tabla **5-2** se presenta a modo resumen la caracterización de los genes ortólogos asociados al rasgo de domesticación de dehiscencia de la vaina, identificados en las tres especies de frijol.

Tabla 5-2.: Caracterización estructural de los genes relacionados al rasgo de domesticación de dehiscencia de la vaina en frijol común, frijol Lima y frijol mungo.

	No. exones	**Longitud total exones**	**Longitud total intrones**	**Longitud del gen**
SHP1				
Frijol común	8	948	6708	7656
Frijol Lima	8	761	7040	7801
Frijol mungo	8	1504	7165	8669
IND				
Frijol común	1	846	0	846
Frijol Lima	1	843	0	843
Frijol mungo	1	842	0	842
ALC				
Frijol común	6	1431	2292	3723
Frijol Lima	6	1348	2237	3585
Frijol mungo	6	1295	2239	3534
FUL				
Frijol común	8	1342	7744	9086
Frijol Lima	8	1303	7589	8892
Frijol mungo	8	1076	6858	7934
NST1				
Frijol común	3	1633	519	2152
Frijol Lima	3	1877	560	2437
Frijol mungo	3	1672	439	2111

5.4. Conclusiones

Se realizó la caracterización de las subregiones genómicas asociadas a cinco genes *SHATERPROOF* (*SHP1*), *INDEHISCENT* (*IND*), *ALCATRAZ* (*ALC*), *FRUITFULL* (*FUL*) y *NST1* (*NAC SECONDARY WALL THICKENING PROMOTING FACTOR 1*) involucrados en el procesos de domesticación especialmente en el rasgo de dehiscencia de la vaina encontrando que para el gen *SHP*, existe una región genómica conservado en cuanto a la orientación del gen de interés y una alta similitud corriente arriba y abajo, adicionalmente a nivel estructural los genes candidatos evaluados presentan una alta consistencia en relación

al número de exónes, y tamaño total del gen.

En el caso de *IND*, este gen se encontró en las tres especies de frijol en un único bloque colineal con un peso de 4.749, sin embargo en frijol lima se encuentra en el reveso complementario. La región genómica evaluada para el gen *ALC*, evidenció una región consistente de un solo bloque colineal, adicionalmente este gen en frijol Lima y frijol común tiene la misma estructura, en cuanto al número de exones y el tamaño total del gen. El gen *FUL*, se encontró estructuralmente altamente conservado, con un tamaño total similar, con el mismo número de exones (8) para las tres especies de frijol evaluadas. Para el gen *NST-1* en frijol lima, común y mungo se identificó una alta similitud estructural con tres exones, en relación con el tamaño del gen, este es similar en las tres especies.

5.5. Materiales y métodos

5.5.1. Obtención de los datos genómicos

Para realizar la comparación de los genomas en especies de leguminosas hermanas se recuperó el archivo de ensamblaje con su respectiva anotación y el proteoma del frijol común y el frijol mungo. Para el frijol Lima se emplearon el ensamblaje, anotación y proteoma generados en la presente investigación. De la base de datos Phytozome se obtuvieron los archivos de frijol común versión 2.1 (https://phytozome.jgi.doe.gov/pz/portal.html) y para frijol mungo versión 6 de la base de datos del NCBI-GenBank
(https://www.ncbi.nlm.nih.gov/nuccore/JJMO00000000.1), junto con los archivos adicionales de anotación de leguminosas disponibles en https://legumeinfo.org/ [17].

5.5.2. Identificación de regiones microsinténicas de genes asociados a domesticación

Para la identificación de los cinco genes candidatos asociados al rasgo de dehiscencia de la vaina se desarrolló una estrategia que consistió en cuatro fases: en la primera fase se emplearon las secuencias de ADN de cada uno de los cinco genes reportados en frjol común que se descargaron de la base de datos Phytozome (https://phytozome.jgi.doe.gov/pz/portal.html). La secuencia de ADN de cada gen se empleó como secuencia problema (o query), en el programa de alineamiento Blastn v 2.2.29 [23] para encontrar secuencias cercanas en el genoma del frijol mungo y frijol Lima. En la segunda fase, los mejores hits encontrados, es decir, las secuencias que compartieron con la secuencia problema uno o más pares de subsecuencias con alta puntuación, fueron empleados para realizar la extracción de las secuencias de acuerdo a las coordenadas genómicas encontradas, mediante scripts desarrollados en el lenguaje

de programación Perl. La región genómica que se extrajo fue de 100 kb corriente arriba y corriente abajo del mejor hit, en cada una de las coordenadas previamente seleccionadas. Adicionalmente, se obtuvo para frijol común a partir de las coordenadas de cada gen la región genómica de 100 Kb.

En la tercera fase se realizó un alineamiento múltiple con las secuencias seleccionadas para cada gen mediante el programa Mauve, el cual identifica y alinea las regiones de colinealidad local, denominadas bloques colineales locales (LCB). Cada bloque es una región de secuencia homóloga compartida por dos o más de los genomas en estudio. En el algoritmo de Mauve se identifican 5 etapas: La primera consiste en encontrar alineamientos locales denominados multi-MUMs, a través de método de semilla y extensión, identificando coincidencias en un subset del genoma que está siendo alineado. La segunda etapa consiste en calcular un árbol guia, donde Mauve explora la información proporcionada por un set de milti-MUMs como una métrica de distancia para construir un árbol filogenético usando un algoritmo de agrupamiento Neighbor Joining. La tercera etapa consiste en la selección de un conjunto de bloques, eliminando los posibles falsos MUMs, de acuerdo a un criterio conocido como peso mínimo. En la cuarta etapa Mauve identifica alineamientos adicionales fuera y dentro de los LCB. Finalmente, en la quinta etapa Mauve realiza un alineamiento progresivo en cada LCB usando el árbol guía.[30].

En la cuarta fase se emplearon los archivos de anotación de cada una de las especies para la caracterización genómica de la región en términos del contenido génico, la orientación de los genes y el grado de conservación de la estructura de los genes en cuanto a número de exones, longitud total de exones y longitud total de intrones.

En la figura **5-16** se detalla las diferentes etapas y herramientas empleadas en la identificación de regiones microsinténicas de la presente investigación.

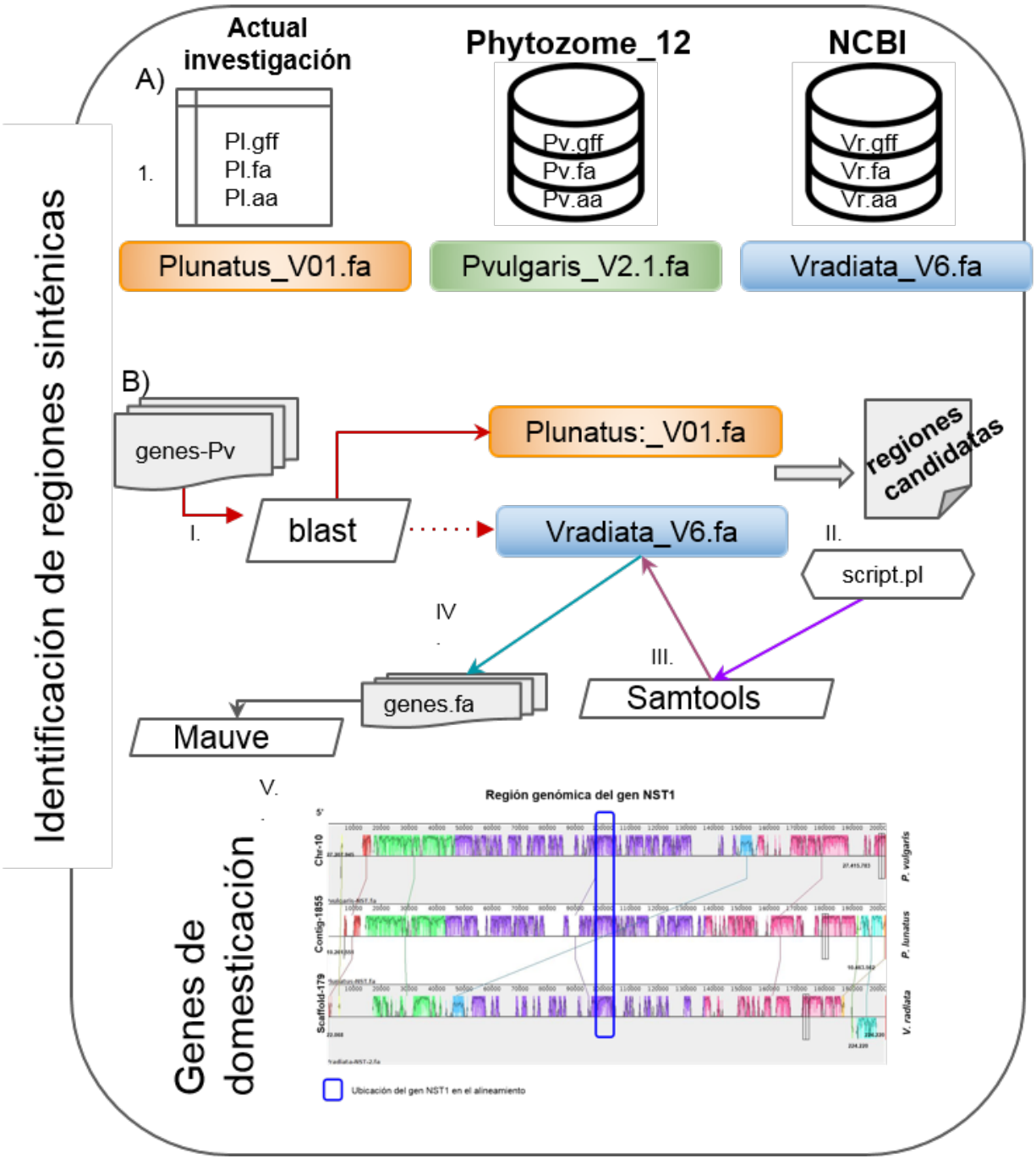


Figura 5-16.: Herramientas para la identificación de regiones sinténicas
A)fase de obtención de los datos. **B)**fase de procesamiento: **I.**Búsqueda por blast II. Identificación de regiones candidatas. **III.**Extracción de sub-regiones genómicas. **IV.**Alineamientos múltiples **V.** Análisis microsinténico.

6. Conclusiones y recomendaciones

En la presente investigación se generó el primer ensamblaje del genoma alta calidad para frijol Lima a nivel de scaffolds, siendo un aporte para la construcción del genoma de referencia. Adicionalmente, el presente ensamblaje da inicio a la posibilidad de realizar estudios a nivel comparativo con las diferentes especies de leguminosas, especialmente con *Phaseolus vulgaris*, que lleven a generar explicaciones en torno a las adaptaciones ecológicas y transformaciones fenotípicas o adaptaciones a la domesticación, desde los posibles cambios genómicos en estas dos especies.

El proceso de ensamblaje *de novo* conlleva una serie de retos computacionales, al intentar reconstruir la secuencia original del genoma de una planta a partir de leturas cortas producidas por tecnologias de secuenciacion de segunda generación, sin embargo la incorporación de lecturas de tercera generación permitió obtener ensamblajes de alta calidad, evidenciado en la alta contigüidad del ensamblaje de frijol Lima con un N50 de 5.5 Mb.

Adicional a la generación del genoma, es necesario realizar su caracterización estructural y funcional. Sin embargo, este proceso conlleva un análisis cuidadoso que permita obtener un alto nivel de curatoría en la anotación, para ello se requieren múltiples rondas de re-anotación, prestando especial cuidado con los procesos de identificación de elementos repetitivos y prediciones *de novo*.

El uso de ensamblajes de transcriptomas de diferentes tejidos aportó el 40 % de los genes anotados en el genoma de frijol Lima, evidenciándose la importancia del uso de datos de RNA-seq en la caracterización de genes en esta especie.

Se caracterizaron cincos genes asociados al rasgo de domesticación de dehiscencia de la vaina, evaluando los genes ortólogos presentes en una sub-región de 200 kb. Se encontró que los genes *SHP1*, *ALC*, *FUL*, *NST1* e *IND* están altamente conservados entre frijol Lima y frijol común en cuanto a su estructura (número de exones, longitud total de exones e introns y tamaño del gen), por ende probablemente conserven su función.

El presente trabajo se constituye en un avance significativo en la generación de recursos genómicos para frijol Lima, y abre nuevas posibilidades en el estudio de la arquitectura genética de rasgos de interés en las leguminosas y en el entendimiento de su evolución.

6.0.1. Aportes en eventos científicos.

Los avances del presente trabajo fueron presentados en:

- Seminario-taller "Arquitectura genética de rasgos complejos en plantas", Facultad de Ciencias Agrarias, Universidad Nacional de Colombia (2018, Bogotá-Colombia), seminario titulado: Avances en el secuenciamiento del genoma de fríjol Lima.

- IV Congreso Colombiano de Biología Computacional y Bioinformática y VIII Conferencia Iberoamericana de Bioinformática (2017, Cali-Colombia) , en la modalidad de póster, con el titulo: Development of genomic resources in Lima beans (*Phaseolus lunatus* L.) for evolutionary studies and future breeding programs. Autores: Tatiana García, Paola Hurtado, Jorge Duitama & María I. Chacón.

A. Anexo: Control de calidad de las librerías genómicas

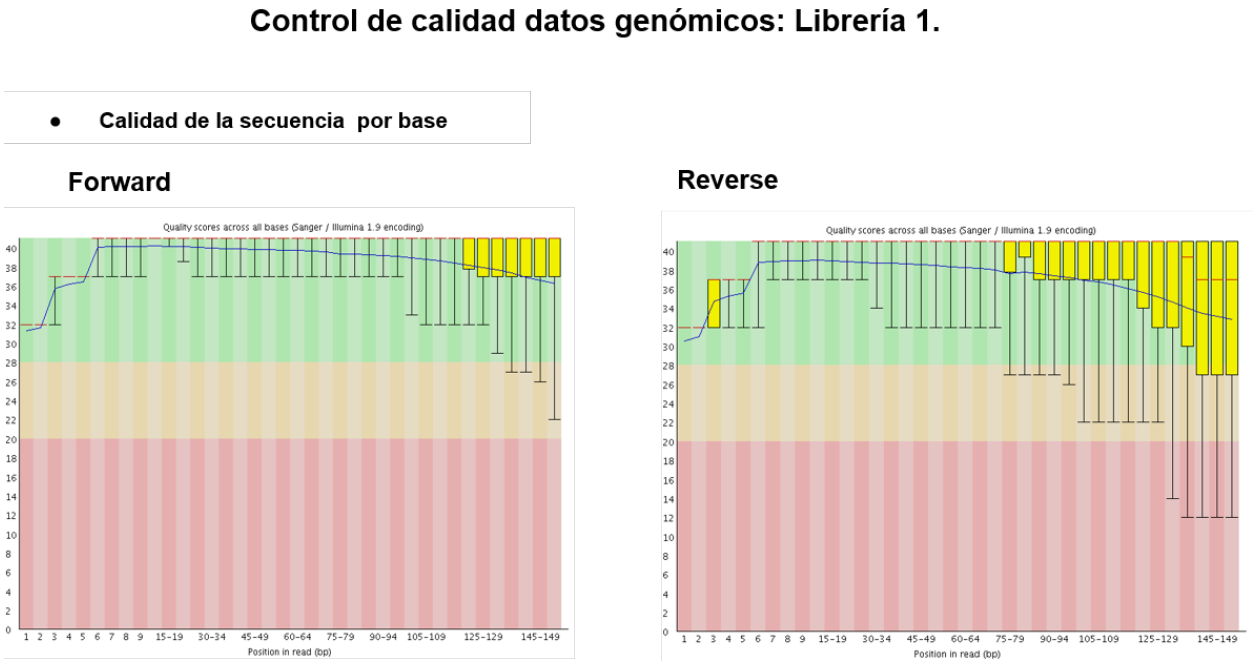
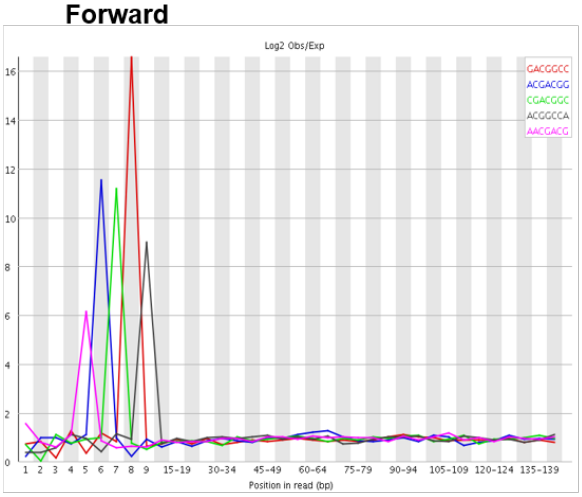


Figura A-1.: Calidad de la secuencia por base de la librería 1.

- **Contenido de Kmers**

Forward



Reverse

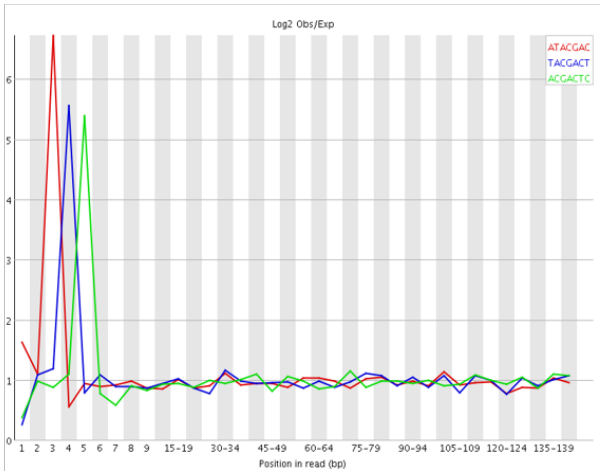
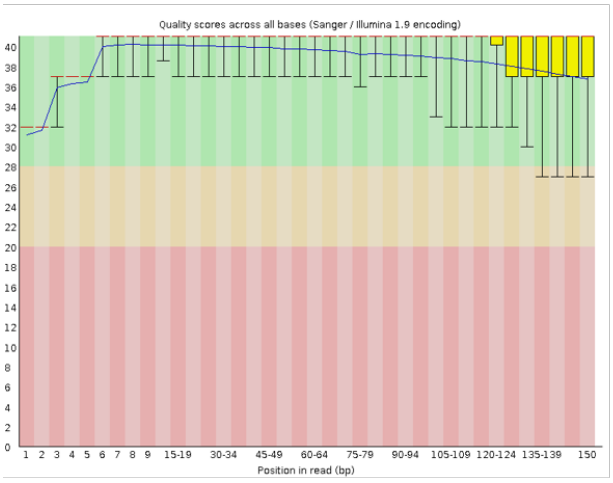


Figura A-2.: Contenido de Kmers de la librería 1.

Control de calidad datos genómicos: Librería 2.

- **Calidad de la secuencia por base**

Forward



Reverse

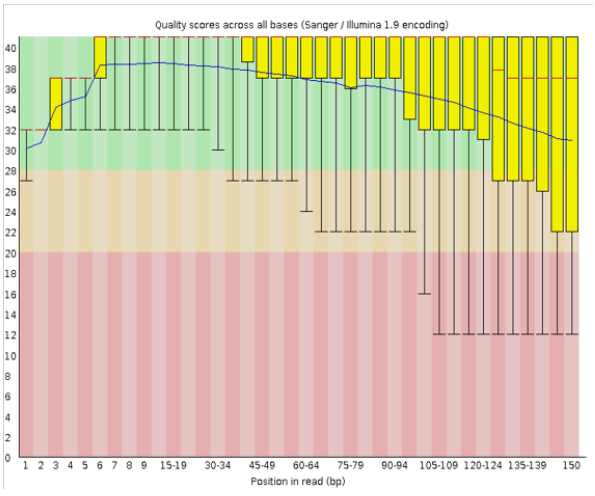


Figura A-3.: Calidad de la secuencia por base de la librería 2, construida con la tecnologia 10X y secuenciadas con la plataforma Illumina.

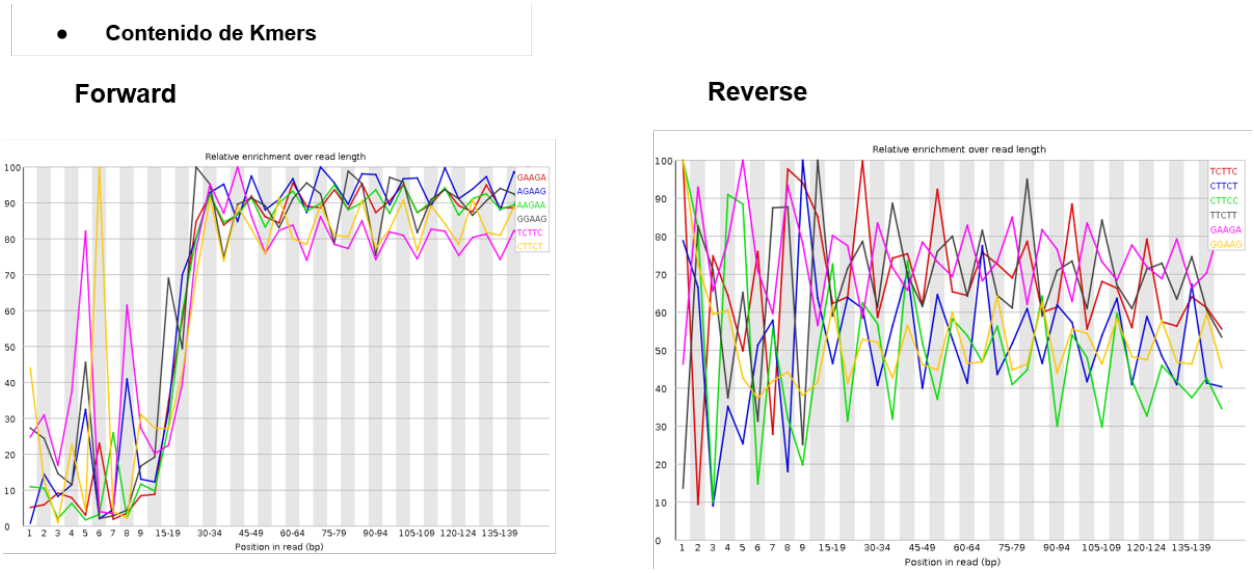


Figura A-4.: Contenido de kmers de la librería 2, construida con la tecnologia 10X y secuenciadas con la plataforma Illumina.

B. Anexo: Control de calidad de las librerías de RNA-seq

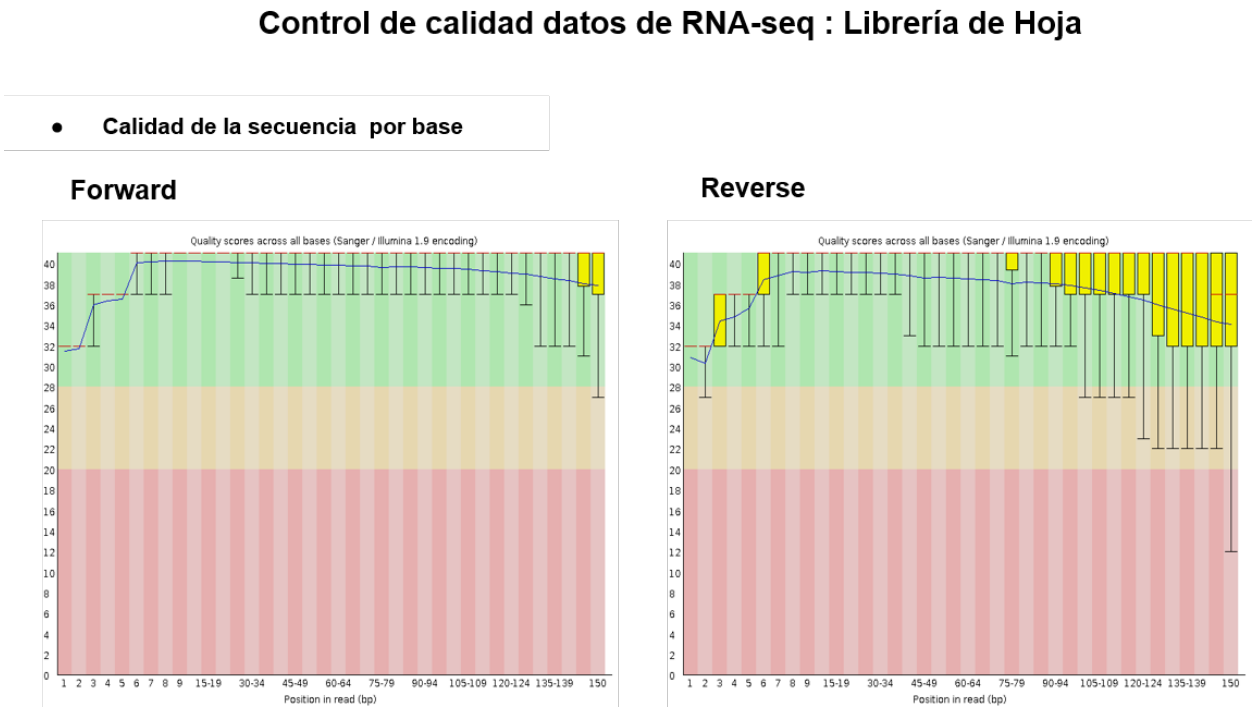


Figura B-1.: Calidad de la secuencia por base de la librería de hoja.

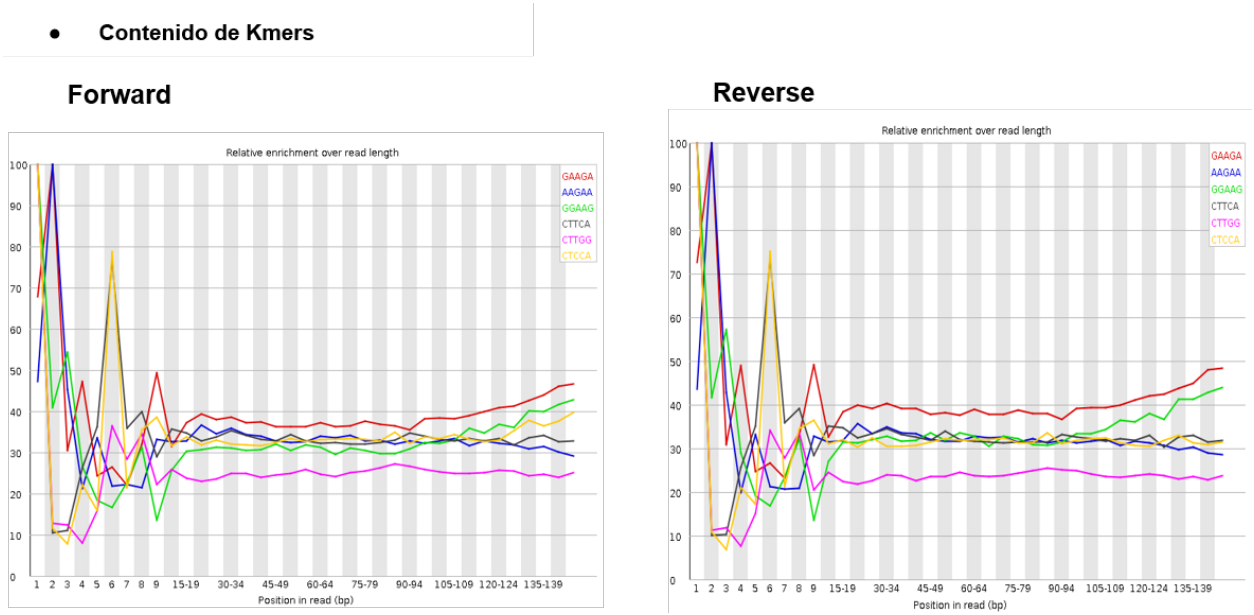


Figura B-2.: Contenido de kmers de la librería de hoja.

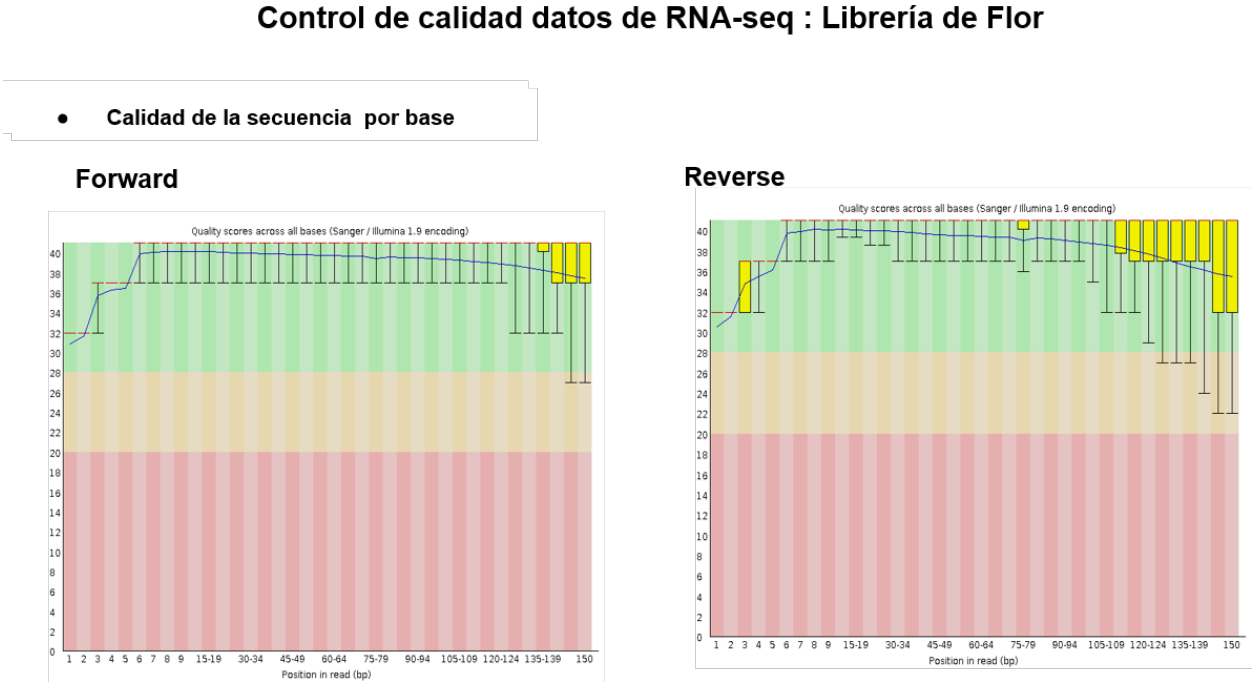


Figura B-3.: Calidad de la secuencia por base de la librería de la flor.

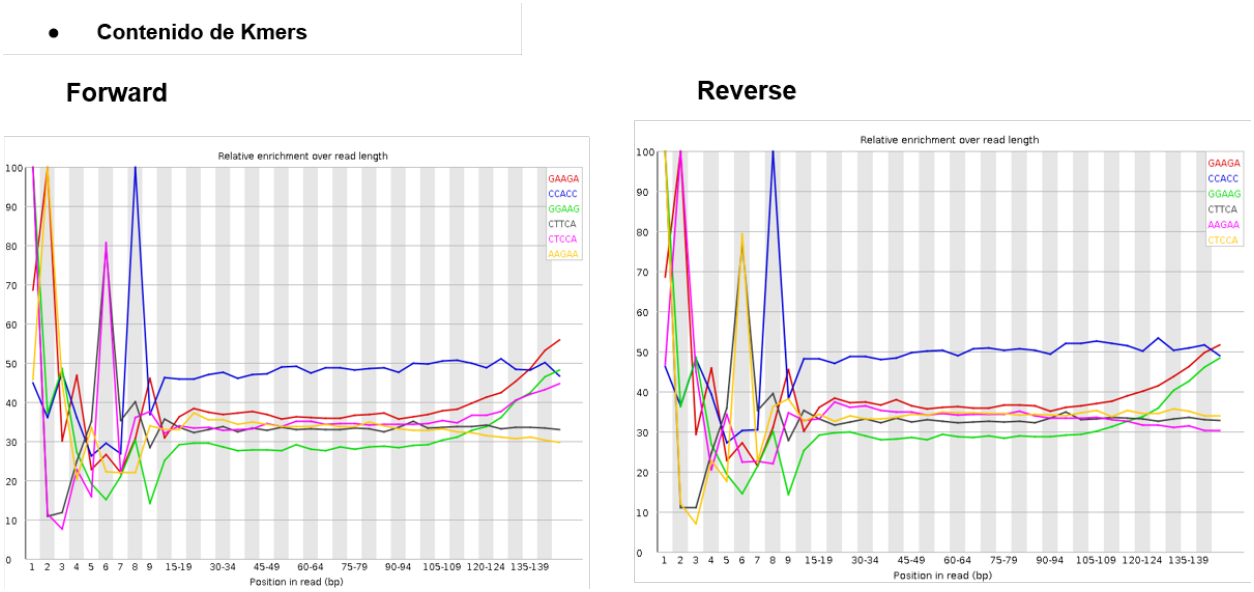


Figura B-4.: Contenido de kmers de la librería de la flor.

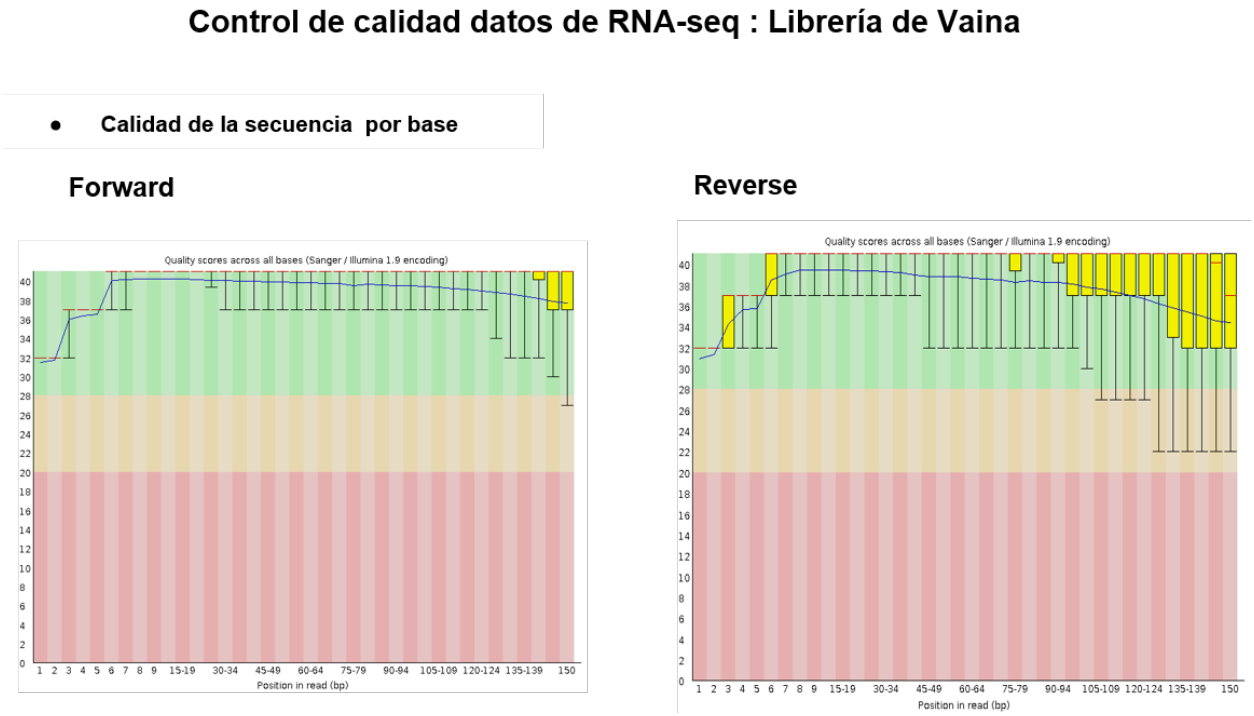


Figura B-5.: Calidad de la secuencia por base de la librería de vaina.

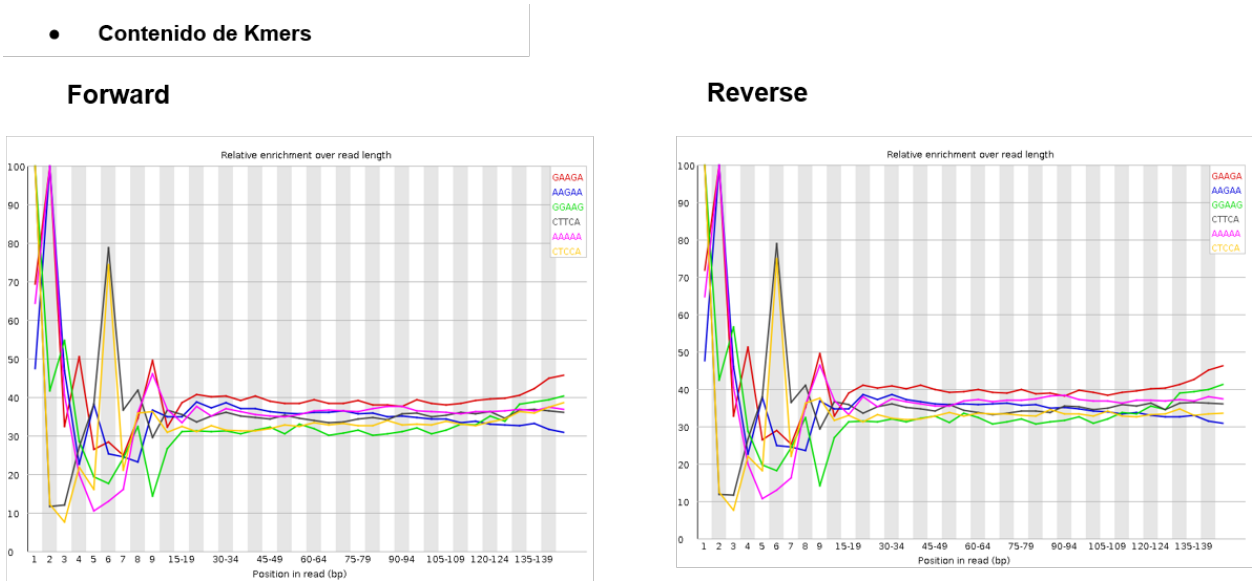


Figura B-6.: Contenido de kmers de la librería de vaina.

C. Anexo: Regiones genómicas asociadas a domesticación

Tabla C-1.: Tamaños de los genes con sus respectivo número de exones e introns en la región subgenómica en *Phaseolus vulgaris* del gen *SHP*

GEN *SHATERPROOF*							
Phaseolus vulgaris							
ID-1	**ID-2**	**coordenada**		**hebra**	**tamaño**	**# exones**	**# introns**
		inicial	**final**				
1-Pv	Phvul.006G168500	27207679	27211382	-	3704	11	10
2-Pv	Phvul.006G168600	27217344	27217953	-	610	1	
3-Pv	Phvul.006G168700	27228090	27230219	+	2130	7	6
4-Pv	Phvul.006G168800	27234042	27240222	-	6181	9	8
5-Pv	Phvul.006G168900	27247085	27248869	+	1785	2	1
6-Pv	Phvul.006G169000	27255067	27258228	+	3162	5	4
7-Pv	Phvul.006G169100	27264424	27267345	+	2922	5	4
8-Pv	Phvul.006G169200	27272887	27276882	+	3996	5	4
9-Pv	Phvul.006G169300	27286715	27289965	+	3251	5	4
10-Pv	Phvul.006G169400	27295400	27297478	+	2079	5	4
11-Pv	Phvul.006G169500	27300758	27303153	+	2396	5	4
12-Pv	Phvul.006G169600	27307945	27315783	-	7839	8	7
13-Pv	Phvul.006G169700	27342279	27349872	-	7594	12	11
14-Pv	Phvul.006G169800	27351512	27351945	+	434	2	1
15-Pv	Phvul.006G169900	27354772	27355308	+	537	1	
16-Pv	Phvul.006G170000	27357329	27358604	-	1276	4	3
17-Pv	Phvul.006G170100	27360116	27363035	-	2920	5	4
18-Pv	Phvul.006G170200	27367777	27369945	+	2169	1	
19-Pv	Phvul.006G170300	27372513	27375328	+	2816	1	
20-Pv	Phvul.006G170500	27382184	27391983	-	9800	24	23
21-Pv	Phvul.006G170400	27379125	27381388	+	2264	1	
22-Pv	Phvul.006G170600	27394206	27398759	+	4554	6	5
23-Pv	Phvul.006G170700	27398890	27403542	-	4653	8	7
24-Pv	Phvul.006G170800	27406154	27408897	+	2744	2	1

Tabla C-2.: Tamaños de los genes con sus respectivo número de exones e introns en la región subgenómica en *Phaseolus lunatus* del gen SHP

GEN *SHATERPROOF*							
Phaseolus lunatus							
ID-1	**ID-2**	**coordenada**		**hebra**	**tamaño**	#	#
		inicial	**final**			**exones**	**introns**
1-Pl	1-Pl	568970	569155	+	185	1	
2-Pl	2-Pl	685041	685523	+	482	1	
3-Pl	3-Pl	683291	683750	+	459	2	1
4-Pl	4-Pl	627007	629178	+	2171	6	5
5-Pl	5-Pl	633039	635660	+	2621	5	4
6-Pl	6-Pl	593651	596461	+	2810	5	4
7-Pl	7-Pl	584942	588558	+	3616	5	4
8-Pl	8-Pl	619073	622661	+	3588	6	5
9-Pl	9-Pl	577692	580278	+	2586	6	5
10-Pl	10-Pl	602146	602409	+	263	1	
11-Pl	11-Pl	612344	612898	+	554	2	1
12-Pl	12-Pl	639777	643620	-	7801	8	6
13-Pl	13-Pl	671726	678999	-	7273	13	12
14-Pl	14-Pl	607492	608661	-	1169	2	1
15-Pl	15-Pl	645170	647911	-	2741	4	3
16-Pl	16-Pl	602507	607329	-	4822	7	6
17-Pl	17-Pl	685333	685775	-	442	2	2
18-Pl	18-Pl	654762	655067	+	305	1	
19-Pl	19-Pl	701443	703251	+	1808	1	
20-Pl	20-Pl	737402	741824	+	4422	6	5
21-Pl	21-Pl	714314	716293	+	1979	1	

Tabla C-3.: Tamaños de los genes con sus respectivo número de exones e introns en la región subgenómica en *Vigna radiata* del gen *SHP*

GEN *SHATERPROOF*							
Vigna radiata							
ID-1	**ID-2**	**coordenada**		**hebra**	**tamaño**	**# exones**	**# introns**
		inicial	**final**				
1-Vr	vradi10g09660	17220880	17221395	-	516	1	
2-Vr	Vradi10g09670	17236992	17238931	+	1940	5	4
3-Vr	Vradi10g09680	17241213	17242276	+	1064	3	2
4-Vr	Vradi10g09690	17247258	17255927	-	8670	8	7
5-Vr	Vradi10g09700	17273195	17281087	-	7893	19	18
6-Vr	Vradi10g09730	17297031	17299694	-	2664	5	4
7-Vr	Vradi10g09710	17281935	17289622	+	7688	7	6
8-Vr	Vradi10g09720	17290628	17291818	+	1190	4	3
9-Vr	Vradi10g09740	17330998	17339938	-	8941	24	23
10-Vr	Vradi10g09750	17344226	17347128	+	2903	6	5
11-Vr	Vradi10g09780	17369631	17371504	-	1874	5	4
12-Vr	Vradi10g09790	17387345	17388280	-	936	1	

Tabla C-4.: Tamaños de los genes con sus respectivo número de exones e introns en la región subgenómica en *Phaseolus vulgaris* del gen *IND*

GEN *IND*							
Phaseolus vulgaris							
ID-1	**ID-2**	**coordenada**		**hebra**	**tamaño**	**# exones**	**# introns**
		inicial	**final**				
1-Pv	Phvul.002G270400	44125511	44131328	+	5818	7	6
2-Pv	Phvul.002G270500	44133980	44137288	+	3309	7	6
3-Pv	Phvul.002G270600	44139105	44139713	-	609	1	
4-Pv	Phvul.002G270700	44142989	44144630	+	1642	3	2
5-Pv	Phvul.002G270800	44150261	44150926	+	666	1	
6-Pv	Phvul.002G270900	44155318	44157803	-	2486	2	1
7-Pv	Phvul.002G0271000	44186987	44188326	+	1340	1	
8-Pv	Phvul.002G271100	44199222	44205529	+	6308	3	3
9-Pv	Phvul.002G271200	44205946	44210215	-	4270	12	11
10-Pv	Phvul.002G271300	44216969	44222195	+	5227	5	4
11-Pv	Phvul.002G271400	44224184	44228271	+	4088	5	4
12-Pv	Phvul.002G271600	44232557	44233756	+	1200	1	
13-Pv	Phvul.002G271500	44229799	44246330	+	16532	18	17
14-Pv	Phvul.002G271700	44247536	44251023	+	3488	4	3
15-Pv	Phvul.002G271800	44251436	44254178	-	2743	2	1
16-Pv	Phvul.002G271900	44257054	44258132	+	1079	3	2
17-Pv	Phvul.002G271904	44265079	44272465	-	7387	20	19

Tabla C-5.: Tamaños de los genes con sus respectivo número de exones e intrones en la región subgenómica en *Phaseolus lunatus* y *Vigna radiata* del gen *IND*

GEN IND							
Phaseolus lunatus							
ID-1	**ID-2**	**coordenada**		**hebra**	**tamaño**	**# exones**	**# intrones**
		inicial	**final**				
1-Pl	1-Pl	1407987	1408166	+	179	2	1
2-Pl	2-Pl	1546049	1547773	+	1724	3	2
3-Pl	3-Pl	1570545	1571882	-	1337	1	
4-Pl	4-Pl	1527096	1527621	+	525	2	1
5-Pl	5-Pl	1505878	1506267	+	389	1	
6-Pl	6-Pl	1733350	1735753	+	2403	3	2
Vigna radiata							
1Vr	Vradi01g11020	21428273	21429363	-	1091	4	3
2Vr	Vradi01g11030	21431871	21433322	+	1452	3	2
3Vr	Vradi01g11040	21453293	21455593	+	2301	4	3
4Vr	Vradi01g11050	21461533	21464122	-	2590	2	1
5Vr	Vradi01g11060	21477739	21481468	+	3730	5	4
6Vr	Vradi01g11070	21496391	21497238	+	848	1	0
7Vr	Vradi01g11080	21506261	21512862	+	6602	4	3
8Vr	Vradi01g11090	21513702	21518276	-	4575	13	12
9Vr	Vradi01g11100	21524556	21529418	-	4863	5	4

Tabla C-6.: Tamaños de los genes con sus respectivo número de exones e introns en la región subgenómica en *Phaseolus vulgaris* del gen *ALC*

GEN ALCATRAZ							
Phaseolus vulgaris							
ID-1	**ID-2**	**coordenada**		**hebra**	**tamaño**	**#**	**#**
		inicial	**final**			**exones**	**intrones**
1-Pv	Phvul.001G022200	1835694	1839208	-	3515	5	4
2-Pv	Phvul.001G022300	1841480	1849477	-	7998	11	10
3-Pv	Phvul.001G022400	1851796	1857416	-	5621	5	4
4-Pv	Phvul.001G022500	1858252	1861618	-	3367	5	4
5-Pv	Phvul.001G022600	1865749	1873339	+	7591	11	10
6-Pv	Phvul.001G022700	1877107	1882530	+	5424	9	8
7-Pv	Phvul.001G022800	1881956	1886682	-	4727	5	4
8-Pv	Phvul.001G022900	1895758	1897514	+	1757	2	1
9-Pv	Phvul.001G023000	1922740	192481	-	2077	2	1
10-Pv	Phvul.001G023100	1926639	1929423	-	2785	4	3
11-Pv	Phvul.001G023200	1931951	1935673	-	3723	6	5
12-Pv	Phvul.001G023300	1942169	1947552	-	5384	6	5
13-Pv	Phvul.001G023400	1959506	1963042	-	3537	4	3
14-Pv	Phvul.001G023500	1965730	1970809	-	5080	9	8
15-Pv	Phvul.001G023600	1974828	1976841	-	2014	3	2
16-Pv	Phvul.001G023700	1997594	1998704	-	1111	1	
17-Pv	Phvul.001G024000	2021245	2023290	+	2046	2	1
18-Pv	Phvul.001G023800	2009098	2013649	+	4552	8	7
19-Pv	Phvul.001G023900	2016678	2016678	-	1638	1	
20-Pv	Phvul.001G024100	2026893	2033562	+	6670	12	11

Tabla C-7.: Tamaños de los genes con sus respectivo número de exones e introns en la región subgeóomica en *Phaseolus lunatus* del gen *ALC*

GEN ALCATRAZ							
Phaseolus lunatus							
ID-1	**ID-2**	**coordenada**		**hebra**	**tamaño**	**# exones**	**# introns**
		inicial	**final**				
1-Pl	1-Pl	15574648	15575625	-	977	1	
2-Pl	2-Pl	15577490	15582247	+	4757	7	6
3-Pl	3-Pl	15584552	15584998	-	446	1	
4-Pl	4-Pl	15597593	15597988	-	395	1	
5-Pl	5-Pl	15623346	15625885	-	2539	3	2
6-Pl	6-Pl	15628081	15633086	-	5005	9	8
7-Pl	7-Pl	15639636	15640910	-	1274	2	1
8-Pl	8-Pl	15646640	15647746	-	1106	4	3
9-Pl	9-Pl	15661150	15666213	-	5063	6	5
10-Pl	10-Pl	15671987	15675572	-	3585	6	5
11-Pl	11-Pl	15677089	15680002	-	2913	3	2
12-Pl	12-Pl	15681785	15683005	-	1220	1	
13-Pl	13-Pl	15701478	15703173	+	1695	2	1
14-Pl	14-Pl	15715591	15721209	+	5618	8	7
15-Pl	15-Pl	15731557	15732459	-	902	4	3
16-Pl	16-Pl	15737396	15744630	-	7234	2	1
17-Pl	17-Pl	15746812	15747169	-	357	2	1
18-Pl	18-Pl	15749489	15750504	+	1015	3	2
19-Pl	19-Pl	15750813	15751486	-	673	2	1
20-Pl	20-Pl	15752743	15753841	-	1098	2	1
21-Pl	21-Pl	15769438	15773445	-	4007	5	4

Tabla C-8.: Tamaños de los genes con sus respectivo número de exones e introns en la región subgenómica en *Vigna radiata* del gen ALC

GEN ALCATRAZ							
Vigna radiata							
ID-1	**ID-2**	**coordenada**		**hebra**	**tamaño**	**#**	**#**
		inicial	**final**			**exones**	**introns**
1-Vr	Vradi06g15010	35076055	35077953	-	1899	4	3
2-Vr	Vradi06g15020	35083229	35085266	-	2038	5	4
3-Vr	Vradi06g15030	35087889	35092301	-	4413	9	8
4-Vr	Vradi06g15040	35096274	35099726	-	3453	3	2
5-Vr	Vradi06g15050	35117056	35122102	-	5047	6	5
6-Vr	Vradi06g15060	35135969	35138176	-	2208	4	3
7-Vr	Vradi06g15070	35143233	35148556	-	5324	5	4
8-Vr	Vradi06g15080	35156689	35159458	-	2770	5	4
9-Vr	Vradi06g15090	35173355	35174655	+	1301	2	1
10-Vr	Vradi06g15110	35190222	35195128	+	4907	8	7
11-Vr	Vradi06g15100	35185949	35190024	-	4076	3	4
12-Vr	Vradi06g15120	35198768	35206242	+	7475	9	8
13-Vr	Vradi06g15130	35209206	35212471	-	3266	5	4
14-Vr	Vradi06g15150	35220072	35227096	-	7025	11	10
15-Vr	Vradi06g15140	35213299	35218673	-	5375	7	6

Tabla C-9.: Tamaños de los genes con sus respectivo número de exones e introns en la región subgenómica en *Phaseolus vulgaris* del gen *FUL*

GEN *FRUITFULL*							
Phaseolus vulgaris							
ID-1	**ID-2**	**coordenada**		**hebra**	**tamaño**	**# exones**	**# introns**
		inicial	**final**				
1-Pv	Phvul.009G202700	30777636	30789099	-	11464	6	5
2-Pv	Phvul.009G202800	30793329	30799660	-	6322	12	11
3-Pv	Phvul.009G202900	30806647	30808901	+	2255	1	
4-Pv	Phvul.009G203000	30811562	30816497	-	4950	14	13
5-Pv	Phvul.009G203100	30820370	30823792	+	3423	6	5
6-Pv	Phvul.009G203200	30824149	30826369	-	2221	6	5
7-Pv	Phvul.009G203300	30837691	30839716	+	2026	4	3
8-Pv	Phvul.009G203400	30856875	30865960	+	9086	8	7
9-Pv	Phvul.009G203500	30877227	30881731	+	4505	3	2
10-Pv	Phvul.009G203600	30885390	30885785	+	396	2	1
11-Pv	Phvul.009G203700	30889694	30902478	-	12785	26	25
12-Pv	Phvul.009G203800	30910393	30911730	-	1338	2	1
13-Pv	Phvul.009G203900	30913612	30917101	+	3490	6	5
14-Pv	Phvul.009G204000	30927854	30930010	+	2157	1	
15-Pv	Phvul.009G204100	30941498	30955830	+	14333	13	12

Tabla C-10.: Tamaños de los genes con sus respectivo número de exones e introns en la región subgenómica en *Phaseolus lunatus* del gen *FUL*

GEN *FRUITFULL*							
Phaseolus lunatus							
ID-1	**ID-2**	**coordenada**		**hebra**	**tamaño**	**#**	**#**
		inicial	**final**			**exones**	**introns**
1-Pl	1-Pl	6409174	6414545	-	5371	11	10
2-Pl	2-Pl	6415796	6416158	-	362	1	
3-Pl	3-Pl	6419824	6422985	+	3161	4	3
4-Pl	4-Pl	6424283	6435848	+	11565	11	10
5-Pl	5-Pl	6438521	6438964	+	443	1	
6-Pl	6-Pl	6452713	6453651	+	938	1	
7-Pl	7-Pl	6465148	6466920	+	1772	3	2
8-Pl	8-Pl	6478491	6491032	-	12541	11	10
9-Pl	9-Pl	6494685	6494976	+	291	2	1
10-Pl	10-Pl	6500134	6504067	+	3933	4	3
11-Pl	11-Pl	6509816	6510367	-	551	2	1
12-Pl	12-Pl	6511795	6521187	+	9392	8	7
13-Pl	13-Pl	6535819	6538102	+	2283	4	3
14-Pl	14-Pl	6550364	6552409	-	2045	6	5
15-Pl	15-Pl	6552628	6556356	+	3728	6	5
16-Pl	16-Pl	6560627	6565570	-	4943	11	10
17-Pl	17-Pl	6570210	6571547	+	1337	1	
18-Pl	18-Pl	6578400	6583033	-	4633	10	9
19-Pl	19-Pl	6588389	6599122	-	10733	5	4
20-Pl	20-Pl	6616404	6621396	-	4992	7	6
21-Pl	21-Pl	6622639	6628452	+	5813	6	5

Tabla C-11.: Tamaños de los genes con sus respectivo número de exones e introns en la región subgenómica en *Vigna radiata* del gen *FUL*

GEN *FRUITFULL*							
Vigna radiata							
ID-1	**ID-2**	**coordenada**		**hebra**	**tamaño**	**# exones**	**# introns**
		inicial	**final**				
1-Vr	1-Vr	24001202	24002500	+	1299	1	
2-Vr	2-Vr	24022048	24023386	-	1339	3	2
3-Vr	3-Vr	24060348	24064282	-	3935	2	1
4-Vr	4-Vr	24080207	24088141	+	7934	8	7
5-Vr	5-Vr	24047223	24051307	+	4085	5	4
6-Vr	6-Vr	24074203	24077304	+	3102	4	3
7-Vr	7-Vr	24097813	24099564	+	1752	3	2
8-Vr	8-Vr	24122915	24134910	-	11996	10	9
9-Vr	9-Vr	24167254	24167788	-	535	2	1

Tabla C-12.: Tamaños de los genes con sus respectivo número de exones e introns en la región subgenómica en *Phaseolus vulgaris* del gen *NST*

GEN *NST*							
Phaseolus vulgaris							
ID-1	**ID-2**	**coordenada**		**hebra**	**tamaño**	**# exones**	**# introns**
		inicial	**final**				
1-Pv	Phvul.010G118400	39776364	39781788	-	5425	4	3
2-Pv	Phvul.010G118500	39795425	39798608	-	3184	5	4
3-Pv	Phvul.010G118600	39807622	39810293	-	2672	7	6
4-Pv	Phvul.010G118700	39859044	39861195	-	2152	3	2
5-Pv	Phvul.010G118800	39887116	39889350	-	2235	1	
6-Pv	Phvul.010G118900	39906332	39906790	-	459	1	
7-Pv	Phvul.010G119000	39914813	39916789	-	1977	1	
8-Pv	Phvul.010G119100	39927363	39928519	-	1157	1	
9-Pv	Phvul.010G119300	39941500	39945893	-	4394	14	13
10-Pv	Phvul.010G119200	39939140	39940586	+	1447	6	5

Tabla C-13.: Tamaños de los genes con sus respectivo número de exones e introns en la región subgenómica en *Phaseolus lunatus* del gen *NST*

GEN NST							
Phaseolus lunatus							
ID-1	**ID-2**	**coordenada**		**hebra**	**tamaño**	**# exones**	**# introns**
		inicial	**final**				
1-Pl	1-Pl	10265004	10265900	+	896	3	2
2-Pl	2-Pl	10271502	10272390	-	888	2	1
3-Pl	3-Pl	10274674	10280893	-	6219	5	4
4-Pl	4-Pl	10283759	10283929	+	170	1	
5-Pl	5-Pl	10290572	10290978	+	406	2	1
6-Pl	6-Pl	10295100	10298278	-	3178	5	4
7-Pl	7-Pl	10306578	10309445	-	2867	5	4
8-Pl	8-Pl	10336849	10337061	-	212	1	
9-Pl	9-Pl	10351811	10352318	-	507	2	1
10-Pl	10-Pl	10361152	10363589	-	2437	3	2
11-Pl	11-Pl	10390181	10392334	-	2153	1	
12-Pl	12-Pl	10398278	10400533	-	2255	1	
13-Pl	13-Pl	10406901	10407440	-	539	1	
14-Pl	14-Pl	10417907	10418056	+	149	1	
15-Pl	15-Pl	10422987	10424340	+	1353	6	5
16-Pl	16-Pl	10425369	10429680	-	4311	14	13
17-Pl	17-Pl	10440846	10452337	+	11491	19	18
18-Pl	18-Pl	10454612	10456695	-	2083	4	3

Tabla C-14.: Tamaños de los genes con sus respectivo número de exones e introns en la región subgenómica en *Vigna radiata* del gen *NST*

GEN *NST*							
Vigna radiata							
ID-1	**ID-2**	**coordenada**		**hebra**	**tamaño**	**# exones**	**# introns**
		inicial	**final**				
1-Vr	1-Vr	25395	27352	-	1957	3	2
2-Vr	2-Vr	28513	32033	-	3520	18	17
3-Vr	3-Vr	38521	51800	+	13279	15	14
4-Vr	4-Vr	57664	61630	+	3966	13	12
5-Vr	5-Vr	61525	63176	+	1651	2	1
6-Vr	6-Vr	63632	65878	+	2246	7	6
7-Vr	7-Vr	73625	74614	+	989	1	
8-Vr	8-Vr	85531	87821	+	2290	1	
9-Vr	9-Vr	122145	124256	+	2111	3	2
10-Vr	10-Vr	92666	94921	+	2255	3	2
11-Vr	11-Vr	154736	155587	+	851	1	
12-Vr	12-Vr	168153	170747	+	2594	5	4
13-Vr	13-Vr	172788	178417	-	5629	3	2
14-Vr	14-Vr	187232	190242	+	3010	3	2
15-Vr	15-Vr	202776	207223	+	4447	5	4

Bibliografía

[1] Andrews, Simon ; others: FastQC: a quality control tool for high throughput sequence data. (2010)

[2] Arumuganathan, Ka ; Earle, ED: Nuclear DNA content of some important plant species. *Plant molecular biology reporter* 9 (1991), Nr. 3, S. 208–218

[3] Altenhoff, Adrian M. ; Gil, Manuel ; Gonnet, Gaston H. ; Dessimoz, Christophe: Inferring hierarchical orthologous groups from orthologous gene pairs. *PLoS One* 8 (2013), Nr. 1, S. e53786

[4] Andueza-Noh, Rubén H ; Serrano-Serrano, Martha L. ; Sánchez, MI C. ; Pino, I S. ; Camacho-Pérez, L ; Coello-Coello, J ; Cortes, J M. ; Debouck, Daniel G. ; Martínez-Castillo, Jaime: Multiple domestications of the Mesoamerican gene pool of lima bean (Phaseolus lunatus L.): evidence from chloroplast DNA sequences. *Genetic Resources and Crop Evolution* 60 (2013), Nr. 3, S. 1069–1086

[5] Abbo, Shahal ; Oss, Ruth P. ; Gopher, Avi ; Saranga, Yehoshua ; Ofner, Itai ; Peleg, Zvi: Plant domestication versus crop evolution: a conceptual framework for cereals and grain legumes. *Trends in Plant Science* 19 (2014), Nr. 6, S. 351–360

[6] Almeida, Cícero ; Pedrosa-Harand, Andrea: High macro-collinearity between lima bean (Phaseolus lunatus L.) and the common bean (P. vulgaris L.) as revealed by comparative cytogenetic mapping. *Theoretical and applied genetics* 126 (2013), Nr. 7, S. 1909–1916

[7] Astudillo-Reyes, Carolina ; Fernandez, Andrea C. ; Cichy, Karen A.: Transcriptome characterization of developing bean (Phaseolus vulgaris L.) pods from two genotypes with contrasting seed Zinc concentrations. *PloS one* 10 (2015), Nr. 9, S. e0137157

[8] Baudoin, Jean-Pierre ; Baudoin, Jean-Pierre ; Rocha, Oscar ; Degreef, Jérôme ; Maquet, Alain ; Guarino, Luigi: *Ecogeography, demography, diversity and conservation of Phaseolus lunatus L. in the central valley of Costa Rica*. Bd. 12. Bioversity International, 2004

[9] Buermans, HPJ ; Den Dunnen, JT: Next generation sequencing technology: advances and applications. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease* 1842 (2014), Nr. 10, S. 1932–1941

[10] Bonifácio, Eliene M. ; Fonsêca, Artur ; Almeida, Cícero ; Santos, Karla G. ; Pedrosa-Harand, Andrea: Comparative cytogenetic mapping between the lima bean (Phaseolus lunatus L.) and the common bean (P. vulgaris L.). *Theoretical and Applied Genetics* 124 (2012), Nr. 8, S. 1513–1520

[11] Bradnam, Keith R. ; Fass, Joseph N. ; Alexandrov, Anton ; Baranay, Paul ; Bechner, Michael ; Birol, Inanç ; Boisvert, Sébastien ; Chapman, Jarrod A. ; Chapuis, Guillaume ; Chikhi, Rayan ; others: Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *GigaScience* 2 (2013), Nr. 1, S. 10

[12] Broughton, William J. ; Hernandez, G ; Blair, M ; Beebe, S ; Gepts, P ; Vanderleyden, Jos: Beans (Phaseolus spp.)–model food legumes. *Plant and soil* 252 (2003), Nr. 1, S. 55–128

[13] Blatch, Gregory L. ; Lässle, Michael: The tetratricopeptide repeat: a structural motif mediating protein-protein interactions. *Bioessays* 21 (1999), Nr. 11, S. 932–939

[14] Bolger, Anthony M. ; Lohse, Marc ; Usadel, Bjoern: Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30 (2014), Nr. 15, S. 2114–2120

[15] Brown, Terry: *Genomas/Genome*. Ed. Médica Panamericana, 2008

[16] Brucher, H: The wild ancestor of Phaseolus vulgaris in South America, S. 185–214

[17] Campbell, Michael S. ; Holt, Carson ; Moore, Barry ; Yandell, Mark: Genome annotation and curation using MAKER and MAKER-P. *Current Protocols in Bioinformatics* (2014), S. 4–11

[18] Cantarel, Brandi L. ; Korf, Ian ; Robb, Sofia M. ; Parra, Genis ; Ross, Eric ; Moore, Barry ; Holt, Carson ; Alvarado, Alejandro S. ; Yandell, Mark: MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome research* 18 (2008), Nr. 1, S. 188–196

[19] Choi, Hong-Kyu ; Mun, Jeong-Hwan ; Kim, Dong-Jin ; Zhu, Hongyan ; Baek, Jong-Min ; Mudge, Joanne ; Roe, Bruce ; Ellis, Noel ; Doyle, Jeff ; Kiss, Gyorgy B. ; others: Estimating genome conservation between crop and model legume species. *Proceedings of the National Academy of Sciences of the United States of America* 101 (2004), Nr. 43, S. 15289–15294

[20] Cannon, Steven B. ; McCombie, W R. ; Sato, S ; Tabata, S ; Denny, R ; Palmer, L ; Katari, M ; Young, ND ; Stacey, G: Evolution and microsynteny of the apyrase gene family in three legume genomes. *Molecular Genetics and Genomics* 270 (2003), Nr. 4, S. 347–361

[21] Conesa, Ana ; Madrigal, Pedro ; Tarazona, Sonia ; Gomez-Cabrero, David ; Cervera, Alejandra ; McPherson, Andrew ; Szcześniak, Michał W. ; Gaffney, Daniel J. ; Elo, Laura L. ; Zhang, Xuegong ; others: A survey of best practices for RNA-seq data analysis. *Genome biology* 17 (2016), Nr. 1, S. 13

[22] Consortium, UniProt: UniProt: a hub for protein information. *Nucleic acids research* 43 (2014), Nr. D1, S. D204–D212

[23] Coordinators, NCBI R.: Database resources of the national center for biotechnology information. *Nucleic acids research* 44 (2016), Nr. Database issue, S. D7

[24] Compeau, Phillip E. ; Pevzner, Pavel A. ; Tesler, Glenn: How to apply de Bruijn graphs to genome assembly. *Nature biotechnology* 29 (2011), Nr. 11, S. 987

[25] Chacón-Sanchez, Maria I.: Darwin y la domesticación de plantas en las Américas: el caso del maíz y el fríjol. *Acta Biológica Colombiana* 14 (2009), Nr. 4s, S. 351–364

[26] Chaney, Lindsay ; Sharp, Aaron R. ; Evans, Carrie R. ; Udall, Joshua A.: Genome mapping in plant comparative genomics. *Trends in plant science* 21 (2016), Nr. 9, S. 770–780

[27] Considine, Michael J. ; Siddique, Kadambot H. ; Foyer, Christine H.: Nature's pulse power: legumes, food security and climate change. *Journal of experimental botany* 68 (2017), Nr. 8, S. 1815–1818

[28] Clément, Yves ; Sarah, Gautier ; Holtz, Yan ; Homa, Felix ; Pointet, Stéphanie ; Contreras, Sandy ; Nabholz, Benoit ; Sabot, François ; Sauné, Laure ; Ardisson, Morgane ; others: Evolutionary forces affecting synonymous variations in plant genomes. *PLoS genetics* 13 (2017), Nr. 5, S. e1006799

[29] Chacón-Sánchez, María I ; Martínez-Castillo, Jaime: Testing domestication scenarios of Lima bean (Phaseolus lunatus L.) in Mesoamerica: insights from genome-wide genetic markers. *Frontiers in Plant Science* 8 (2017), S. 1551

[30] Darling, Aaron C. ; Mau, Bob ; Blattner, Frederick R. ; Perna, Nicole T.: Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome research* 14 (2004), Nr. 7, S. 1394–1403

[31] Delgado-Salinas, Alfonso ; Bibler, Ryan ; Lavin, Matt: Phylogeny of the genus Phaseolus (Leguminosae): a recent diversification in an ancient landscape. *Systematic Botany* 31 (2006), Nr. 4, S. 779–791

[32] Duitama, Jorge ; Silva, Alexander ; Sanabria, Yamid ; Cruz, Daniel F. ; Quintero, Constanza ; Ballen, Carolina ; Lorieux, Mathias ; Scheffler, Brian ; Farmer, Andrew

; Torres, Edgar ; others: Whole genome sequencing of elite rice cultivars as a comprehensive information resource for marker assisted selection. *PLoS One* 10 (2015), Nr. 4, S. e0124617

[33] Delgado-Salinas, Alfonso ; Turley, Tom ; Richman, Adam ; Lavin, Matt: Phylogenetic analysis of the cultivated and wild species of Phaseolus (Fabaceae). *Systematic Botany* (1999), S. 438–460

[34] Du, Huilong ; Yu, Ying ; Ma, Yanfei ; Gao, Qiang ; Cao, Yinghao ; Chen, Zhuo ; Ma, Bin ; Qi, Ming ; Li, Yan ; Zhao, Xianfeng ; others: Sequencing and de novo assembly of a near complete indica rice genome. *Nature communications* 8 (2017), S. 15324

[35] Estornell, Leandro H. ; Agustí, Javier ; Merelo, Paz ; Talón, Manuel ; Tadeo, Francisco R.: Elucidating mechanisms underlying organ abscission. *Plant Science* 199 (2013), S. 48–60

[36] Esposito, Alfonso ; Colantuono, Chiara ; Ruggieri, Valentino ; Chiusano, Maria L.: Bioinformatics for agriculture in the next-generation sequencing era. *Chemical and Biological Technologies in Agriculture* 3 (2016), Nr. 1, S. 9

[37] El-Metwally, Sara ; Ouda, Osama M. ; Helmy, Mohamed: *Next generation sequencing technologies and challenges in sequence assembly*. Bd. 7. Springer Science & Business, 2014

[38] Endicott, Jane A. ; Noble, Martin E. ; Johnson, Louise N.: The structural basis for control of eukaryotic protein kinases. *Annual review of biochemistry* 81 (2012), S. 587–613

[39] Ekblom, Robert ; Wolf, Jochen B.: A field guide to whole-genome sequencing, assembly and annotation. *Evolutionary applications* 7 (2014), Nr. 9, S. 1026–1042

[40] Fuller, Dorian Q. ; Allaby, Robin: Seed dispersal and crop domestication: shattering, germination and seasonality in evolution under cultivation. *Annual Plant Reviews Volume 38: Fruit Development and Seed Dispersal* (2009), S. 238–295

[41] Fonseca, Rute R. ; Albrechtsen, Anders ; Themudo, Gonçalo E. ; Ramos-Madrigal, Jazmín ; Sibbesen, Jonas A. ; Maretty, Lasse ; Zepeda-Mendoza, M L. ; Campos, Paula F. ; Heller, Rasmus ; Pereira, Ricardo J.: Next-generation biology: sequencing and data analysis approaches for non-model organisms. *Marine genomics* 30 (2016), S. 3–13

[42] Ferrándiz, Cristina ; Liljegren, Sarah J. ; Yanofsky, Martin F.: Negative regulation of the SHATTERPROOF genes by FRUITFULL during Arabidopsis fruit development. *Science* 289 (2000), Nr. 5478, S. 436–438

[43] Garg, Rohini ; others: Transcriptome analyses in legumes: A resource for functional genomics. *The Plant Genome* 6 (2013), Nr. 3

[44] Gonzales, Michael D. ; Archuleta, Eric ; Farmer, Andrew ; Gajendran, Kamal ; Grant, David ; Shoemaker, Randy ; Beavis, William D. ; Waugh, Mark E.: The Legume Information System (LIS): an integrated information resource for comparative legume biology. *Nucleic Acids Research* 33 (2005), Nr. suppl_1, S. D660–D665

[45] Gepts, Paul ; Beavis, William D. ; Brummer, E C. ; Shoemaker, Randy C. ; Stalker, H T. ; Weeden, Norman F. ; Young, Nevin D. Legumes as a model plant family. Genomics for food and feed report of the cross-legume advances through genomics conference. 2005

[46] Genomics 10x. 10x Genomics, tipo @ONLINE. 2018

[47] Gepts, Paul: The contribution of genetic and genomic approaches to plant domestication studies. *Current opinion in plant biology* 18 (2014), S. 51–59

[48] Grabherr, Manfred G. ; Haas, Brian J. ; Yassour, Moran ; Levin, Joshua Z. ; Thompson, Dawn A. ; Amit, Ido ; Adiconis, Xian ; Fan, Lin ; Raychowdhury, Raktima ; Zeng, Qiandong ; others: Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature biotechnology* 29 (2011), Nr. 7, S. 644

[49] Gioia, Tania ; Logozzo, Giuseppina ; Kami, James ; Spagnoletti Zeuli, Pierluigi ; Gepts, Paul: Identification and characterization of a homologue to the Arabidopsis INDEHISCENT gene in common bean. *Journal of Heredity* 104 (2012), Nr. 2, S. 273–286

[50] Goodwin, Sara ; McPherson, John D. ; McCombie, W R.: Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics* 17 (2016), Nr. 6, S. 333

[51] Goodstein, David M. ; Shu, Shengqiang ; Howson, Russell ; Neupane, Rochak ; Hayes, Richard D. ; Fazo, Joni ; Mitros, Therese ; Dirks, William ; Hellsten, Uffe ; Putnam, Nicholas ; others: Phytozome: a comparative platform for green plant genomics. *Nucleic acids research* 40 (2011), Nr. D1, S. D1178–D1186

[52] Gebhardt, Christiane ; Schmidt, Renate ; Schneider, Katharina: Plant genome analysis: the state of the art. *International review of cytology* 247 (2005), S. 223–284

[53] Harlan, Jack R. ; others: *Crops and man.*. American Society of Agronomy, 1992 (Ed. 2)

[54] Heather, James M. ; Chain, Benjamin: The sequence of sequencers: the history of sequencing DNA. *Genomics* 107 (2016), Nr. 1, S. 1–8

[55] Heslop-Harrison, J S. ; Schwarzacher, Trude: Domestication, genomics and the future for banana. *Annals of botany* 100 (2007), Nr. 5, S. 1073–1084

[56] Holland, Bridie ; Widdowson, Elsie M. ; Unwin, ID ; Buss, DH: *Vegetables, herbs and spices: Fifth supplement to McCance and Widdowson's The Composition of Foods*. Bd. 5. Royal Society of Chemistry, 1991

[57] Honaas, Loren A. ; Wafula, Eric K. ; Wickett, Norman J. ; Der, Joshua P. ; Zhang, Yeting ; Edger, Patrick P. ; Altman, Naomi S. ; Pires, J C. ; Leebens-Mack, James H. ; others: Selecting superior de novo transcriptome assemblies: lessons learned by leveraging the best plant genome. *PLoS One* 11 (2016), Nr. 1, S. e0146062

[58] Ibeawuchi, II: Landrace legumes: Synopsis of the culture, importance, potentials and roles in agricultural production systems. *Journal of Biological Sciences* 7 (2007), Nr. 3, S. 464–474

[59] Jones, DB ; Gersdorff, CEF ; Johns, CO ; Finks, AJ: The Proteins of the Lima Bean, Phaseolus lunatus. *Journal of Biological Chemistry* 53 (1922), Nr. 2, S. 231–240

[60] Kuzniar, Arnold ; Ham, Roeland C. ; Pongor, Sándor ; Leunissen, Jack A.: The quest for orthologs: finding the corresponding gene across genomes. *Trends in Genetics* 24 (2008), Nr. 11, S. 539–551

[61] Kang, Yang J. ; Kim, Sue K. ; Kim, Moon Y. ; Lestari, Puji ; Kim, Kil H. ; Ha, Bo-Keun ; Jun, Tae H. ; Hwang, Won J. ; Lee, Taeyoung ; Lee, Jayern ; others: Genome sequence of mungbean and insights into evolution within Vigna species. *Nature communications* 5 (2014), S. ncomms6443

[62] Kalavacharla, Venu ; Liu, Zhanji ; Meyers, Blake C. ; Thimmapuram, Jyothi ; Melmaiee, Kalpalatha: Identification and analysis of common bean (Phaseolus vulgaris L.) transcriptomes by massively parallel pyrosequencing. *BMC plant biology* 11 (2011), Nr. 1, S. 135

[63] Kim, Daehwan ; Langmead, Ben ; Salzberg, Steven L.: HISAT: a fast spliced aligner with low memory requirements. *Nature methods* 12 (2015), Nr. 4, S. 357

[64] Koonin, Eugene V.: Orthologs, paralogs, and evolutionary genomics. *Annu. Rev. Genet.* 39 (2005), S. 309–338

[65] König, Stefanie ; Romoth, Lars ; Stanke, Mario: Comparative Genome Annotation, S. 189–212

[66] Koinange, Epimaki M. ; Singh, Shree P. ; Gepts, Paul: Genetic control of the domestication syndrome in common bean. *Crop Science* 36 (1996), Nr. 4, S. 1037–1045

[67] Koren, Sergey ; Walenz, Brian P. ; Berlin, Konstantin ; Miller, Jason R. ; Bergman, Nicholas H. ; Phillippy, Adam M.: Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome research* 27 (2017), Nr. 5, S. 722–736

[68] Lucas, Joseph M. ; Crollius, Hugues R.: High precision detection of conserved segments from synteny blocks. *PloS one* 12 (2017), Nr. 7, S. e0180198

[69] Li, Fengqi ; Cao, Depan ; Liu, Yang ; Yang, Ting ; Wang, Guirong: Transcriptome sequencing of lima bean (Phaseolus lunatus) to identify putative positive selection in Phaseolus and legumes. *International journal of molecular sciences* 16 (2015), Nr. 7, S. 15172–15187

[70] Liljegren, Sarah J. ; Ditta, Gary S. ; Eshed, Yuval ; Savidge, Beth ; Bowman, John L. ; Yanofsky, Martin F.: SHATTERPROOF MADS-box genes control seed dispersal in Arabidopsis. *Nature* 404 (2000), Nr. 6779, S. 766

[71] Lee, Hayan ; Gurtowski, James ; Yoo, Shinjae ; Nattestad, Maria ; Marcus, Shoshana ; Goodwin, Sara ; McCombie, W R. ; Schatz, Michael: Third-generation sequencing and the future of genomics. *BioRxiv* (2016), S. 048603

[72] Luo, Ruibang ; Liu, Binghang ; Xie, Yinlong ; Li, Zhenyu ; Huang, Weihua ; Yuan, Jianying ; He, Guangzhu ; Chen, Yanxiang ; Pan, Qi ; Liu, Yunjie ; others: SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience* 1 (2012), Nr. 1, S. 18

[73] Liljegren, Sarah J. ; Roeder, Adrienne H. ; Kempin, Sherry A. ; Gremski, Kristina ; Østergaard, Lars ; Guimil, Sonia ; Reyes, Daengnoy K. ; Yanofsky, Martin F.: Control of fruit patterning in Arabidopsis by INDEHISCENT. *Cell* 116 (2004), Nr. 6, S. 843–853

[74] Langmead, Ben ; Salzberg, Steven L.: Fast gapped-read alignment with Bowtie 2. *Nature methods* 9 (2012), Nr. 4, S. 357

[75] Manna, Sam: An overview of pentatricopeptide repeat proteins and their applications. *Biochimie* 113 (2015), S. 93–99

[76] Mehrotra, Shweta ; Goyal, Vinod: Repetitive sequences in plant nuclear DNA: types, distribution, evolution and function. *Genomics, proteomics & bioinformatics* 12 (2014), Nr. 4, S. 164–171

[77] Magrini, Vincent ; Gao, Xin ; Rosa, Bruce A. ; McGrath, Sean ; Zhang, Xu ; Hallsworth-Pepin, Kymberlie ; Martin, John ; Hawdon, John ; Wilson, Richard K. ; Mitreva, Makedonka: Improving eukaryotic genome annotation using single molecule mRNA sequencing. *BMC genomics* 19 (2018), Nr. 1, S. 172

[78] Moreton, Joanna ; Izquierdo, Abril ; Emes, Richard D.: Assembly, assessment, and availability of de novo generated eukaryotic transcriptomes. *Frontiers in genetics* 6 (2016), S. 361

[79] Michael, Todd P. ; Jupe, Florian ; Bemm, Felix ; Motley, Stanley T. ; Sandoval, Justin P. ; Loudet, Olivier ; Weigel, Detlef ; Ecker, Joseph R.: High contiguity Arabidopsis thaliana genome assembly with a single nanopore flow cell. *bioRxiv* (2017), S. 149997

[80] Miller, Jason R. ; Koren, Sergey ; Sutton, Granger: Assembly algorithms for next-generation sequencing data. *Genomics* 95 (2010), Nr. 6, S. 315–327

[81] Mitsuda, Nobutaka ; Ohme-Takagi, Masaru: NAC transcription factors NST1 and NST3 regulate pod shattering in a partially redundant manner by promoting secondary wall formation after the establishment of tissue identity. *The Plant Journal* 56 (2008), Nr. 5, S. 768–778

[82] Mercado-Ruaro, Pedro ; Delgado-Salinas, Alfonso: Karyotypic studies on species of Phaseolus (Fabaceae: Phaseolinae). *American Journal of Botany* 85 (1998), Nr. 1, S. 1–1

[83] Martin, Jeffrey A. ; Wang, Zhong: Next-generation transcriptome assembly. *Nature Reviews Genetics* 12 (2011), Nr. 10, S. 671

[84] Niedringhaus, Thomas P. ; Milanova, Denitsa ; Kerby, Matthew B. ; Snyder, Michael P. ; Barron, Annelise E.: Landscape of next-generation sequencing technologies. *Analytical chemistry* 83 (2011), Nr. 12, S. 4327–4341

[85] Nichio, Bruno T. ; Marchaukoski, Jeroniza N. ; Raittz, Roberto T.: New Tools in Orthology Analysis: A Brief Review of Promising Perspectives. *Frontiers in genetics* 8 (2017)

[86] O'rourke, Jamie A. ; Bolon, Yung-Tsi ; Bucciarelli, Bruna ; Vance, Carroll P.: Legume genomics: understanding biology through DNA and RNA sequencing. *Annals of botany* 113 (2014), Nr. 7, S. 1107–1120

[87] Ogawa, Mikihiro ; Kay, Pippa ; Wilson, Sarah ; Swain, Stephen M.: ARABIDOPSIS DEHISCENCE ZONE POLYGALACTURONASE1 (ADPG1), ADPG2, and QUARTET2 are polygalacturonases required for cell separation during reproductive development in Arabidopsis. *The Plant Cell* 21 (2009), Nr. 1, S. 216–233

[88] Osorno, Juan M. ; McClean, Phillip E.: Common bean genomics and its applications in breeding programs, S. 185–206

[89] Ozsolak, Fatih ; Milos, Patrice M.: RNA sequencing: advances, challenges and opportunities. *Nature reviews genetics* 12 (2011), Nr. 2, S. 87

[90] Perea, Claudia ; De La Hoz, Juan F. ; Cruz, Daniel F. ; Lobaton, Juan D. ; Izquierdo, Paulo ; Quintero, Juan C. ; Raatz, Bodo ; Duitama, Jorge: Bioinformatic analysis of genotype by sequencing (GBS) data with NGSEP. *BMC genomics* 17 (2016), Nr. 5, S. 498

[91] Pevsner, Jonathan: *Bioinformatics and functional genomics*. John Wiley & Sons, 2015

[92] Rhoads, Anthony ; Au, Kin F.: PacBio sequencing and its applications. *Genomics, proteomics & bioinformatics* 13 (2015), Nr. 5, S. 278–289

[93] Rajani, Sarojam ; Sundaresan, Venkatesan: The Arabidopsis myc/bHLH gene ALCATRAZ enables cell separation in fruit dehiscence. *Current Biology* 11 (2001), Nr. 24, S. 1914–1922

[94] Reuter, Jason A. ; Spacek, Damek V. ; Snyder, Michael P.: High-throughput sequencing technologies. *Molecular cell* 58 (2015), Nr. 4, S. 586–597

[95] Salgado, A G. ; Gepts, P ; Debouck, Daniel G.: Evidence for two gene pools of the Lima bean, Phaseolus lunatus L., in the Americas. *Genetic Resources and Crop Evolution* 42 (1995), Nr. 1, S. 15–28

[96] Sato, Shusei ; Isobe, Sachiko ; Tabata, Satoshi: Structural analyses of the genomes in legumes. *Current opinion in plant biology* 13 (2010), Nr. 2, S. 146–152

[97] Stanke, Mario ; Keller, Oliver ; Gunduz, Irfan ; Hayes, Alec ; Waack, Stephan ; Morgenstern, Burkhard: AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic acids research* 34 (2006), Nr. suppl_2, S. W435–W439

[98] Schmutz, Jeremy ; McClean, Phillip E. ; Mamidi, Sujan ; Wu, G A. ; Cannon, Steven B. ; Grimwood, Jane ; Jenkins, Jerry ; Shu, Shengqiang ; Song, Qijian ; Chavarro, Carolina ; others: A reference genome for common bean and genome-wide analysis of dual domestications. *Nature genetics* 46 (2014), Nr. 7, S. 707

[99] Sohn, Jang-il ; Nam, Jin-Wu: The present and future of de novo whole-genome assembly. *Briefings in bioinformatics* 19 (2016), Nr. 1, S. 23–40

[100] Serrano-Serrano, Martha L. ; Andueza-Noh, Rubén H ; Martínez-Castillo, Jaime ; Debouck, Daniel G. ; Chacón, S ; María, I ; others: Evolution and domestication of lima bean in Mexico: Evidence from ribosomal DNA. *Crop Science* 52 (2012), Nr. 4, S. 1698–1712

[101] Stein, Joshua C. ; Yu, Yeisoo ; Copetti, Dario ; Zwickl, Derrick J. ; Zhang, Li ; Zhang, Chengjun ; Chougule, Kapeel ; Gao, Dongying ; Iwata, Aiko ; Goicoechea, Jose L. ;

others: Genomes of 13 domesticated and wild rice relatives highlight genetic conservation, turnover and innovation across the genus Oryza. *Nature Genetics* (2018), S. 1

[102] Tekaia, Fredj: Inferring orthologs: open questions and perspectives. *Genomics Insights* 9 (2016), S. GEI–S37925

[103] Tempel, Sébastien: Using and understanding RepeatMasker, S. 29–51

[104] Vergara, Ismael A. ; Chen, Nansheng: Large synteny blocks revealed between Caenorhabditis elegans and Caenorhabditis briggsae genomes using OrthoCluster. *BMC genomics* 11 (2010), Nr. 1, S. 516

[105] Vlasova, Anna ; Capella-Gutiérrez, Salvador ; Rendón-Anaya, Martha ; Hernández-Oñate, Miguel ; Minoche, André E ; Erb, Ionas ; Câmara, Francisco ; Prieto-Barja, Pablo ; Corvelo, André ; Sanseverino, Walter ; others: Genome and transcriptome analysis of the Mesoamerican common bean and the role of gene duplications in establishing tissue and temporal specialization of genes. *Genome biology* 17 (2016), Nr. 1, S. 32

[106] Wang, Sufang ; Gribskov, Michael: Comprehensive evaluation of de novo transcriptome assembly programs and their effects on differential gene expression analysis. *Bioinformatics* 33 (2017), Nr. 3, S. 327–333

[107] Wang, Zhong ; Gerstein, Mark ; Snyder, Michael: RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews genetics* 10 (2009), Nr. 1, S. 57

[108] Ward, Judson A. ; Ponnala, Lalit ; Weber, Courtney A.: Strategies for transcriptome analysis in nonmodel plants. *American Journal of Botany* 99 (2012), Nr. 2, S. 267–276

[109] Waterhouse, Robert M. ; Seppey, Mathieu ; Simão, Felipe A. ; Manni, Mosè ; Ioannidis, Panagiotis ; Klioutchnikov, Guennadi ; Kriventseva, Evgenia V. ; Zdobnov, Evgeny M.: BUSCO applications from quality assessments to gene prediction and phylogenomics. *Molecular biology and evolution* (2017)

[110] Wu, Jing ; Wang, Lanfen ; Li, Long ; Wang, Shumin: De novo assembly of the common bean transcriptome using short reads for the discovery of drought-responsive genes. *PLoS One* 9 (2014), Nr. 10, S. e109262

[111] Yandell, Mark ; Ence, Daniel: A beginner's guide to eukaryotic genome annotation. *Nature Reviews Genetics* 13 (2012), Nr. 5, S. 329

[112] Yang, In S. ; Kim, Sangwoo: Analysis of whole transcriptome sequencing data: workflow and software. *Genomics & informatics* 13 (2015), Nr. 4, S. 119–125

[113] Zhao, Tao ; Schranz, M E.: Network approaches for plant phylogenomic synteny analysis. *Current opinion in plant biology* 36 (2017), S. 129–134

[114] Zdobnov, Evgeny M. ; Tegenfeldt, Fredrik ; Kuznetsov, Dmitry ; Waterhouse, Robert M. ; Simao, Felipe A. ; Ioannidis, Panagiotis ; Seppey, Mathieu ; Loetscher, Alexis ; Kriventseva, Evgenia V.: OrthoDB v9. 1: cataloging evolutionary and functional annotations for animal, fungal, plant, archaeal, bacterial and viral orthologs. *Nucleic acids research* 45 (2016), Nr. D1, S. D744–D749